TAMPERE UNIVERSITY OF TECHNOLOGY

MIKKO ROININEN

AUDIOVISUAL SENSING AND CONTEXT RECOGNITION FOR MOBILE DEVICES

Master of Science Thesis

Examiners: Prof. Moncef Gabbouj and Dr. Esin Guldogan
Examiners and topic approved in the Faculty of Computing and Electrical Engineering Council meeting 07.04.2010

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan koulutusohjelma
MIKKO ROININEN: Audiovisuaalinen aistiminen ja käyttöympäristön tunnistus mobiililaitteilla
Diplomityö, 73 sivua
Marraskuu 2010
Pääaine: Signaalinkäsittely
Tarkastajat: Professori Moncef Gabbouj, TkT Esin Guldogan
Avainsanat: Sisältöpohjainen videoanalyysi, koneoppiminen, ohjattu luokitus, tukivektorikone, multimodaalisuus, multimodaalinen fuusio, geneettiset algoritmit

Laajojen videotietokantojen täysimittainen hyödyntäminen vaatii tehokkaita automaattisia analyysimenetelmiä. Sisältöpohjainen semanttinen videoanalyysi mahdollistaa automaattisen sisällön merkityksen tulkinnan. Sisältöpohjainen videoanalyysi ajatellaan usein koneoppimisongelmana, jossa ohjattua luokitusta sovelletaan videon avainkehyksistä irroitetuihin piirteisiin. Useimmiten videotiedostot sisältävät kuvan lisäksi ääniraidan, jonka täydentävä informaatio voi merkittävästi helpottaa semanttista analyysia. Visuaalisen ja akustisen informaation yhdistämistä on tästä syystä tutkittu alan kirjallisuudessa. Tällainen multimodaalinen fuusio on hankalaa johtuen rinnakkaisen prosessoinnin ajoituksen, synkronoinnin ja yksittäisten informaatiovirtojen soveltuvuuden vaihtelun ongelmista. Käyttöympäristön tunnistus videotallenteesta tarjoaa mielenkiintoisia mahdollisuuksia mobiililaitteiden käyttöympäristötietoisuuteen, kuten automaattiseen käyttöprofiilin valintaan ja tilannetietoiseen palveluiden muokkaukseen.

Tässä diplomityössä esitellään järjestelmätoteutus sisältöpohjaiseen videon tallennusympäristön tunnistukseen hyödyntäen ääntä ja kuvaa. Visuaalinen tunnistus on toteutettu tukivektorikoneilla kuudesta kuvapiirteestä. Äänipohjainen tunnistus tapahtuu valmiilla ääniympäristöntunnistusjärjestelmällä. Järjestelmä tarjoaa ääni- ja kuvatunnistuksen yhdistämiseen tukivektorikonepohjaisen sekä viisi sääntöpohjaista menetelmää. Sääntöpohjaisissa yhdistämismenetelmissä erilliset luokittelijat painotetaan painokertoimilla, jotka on optimoitu geneettisellä algoritmilla. Järjestelmän ja yhdistämismenetelmien toimivuutta on testattu videotietokannalla, joka on kuvattu 21 arkiympäristössä. Koetulosten perusteella multimodaalinen tunnistus toimii erillisiä ääni- ja kuvamenetelmiä paremmin. Paras tunnistustulos saatiin aikaan tukivektorikonepohjaisella yhdistämisellä.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Information Technology
MIKKO ROININEN: Audiovisual Sensing and Context Recognition for Mobile Devices
Master of Science Thesis, 73 pages
November 2010
Major: Signal Processing
Examiners: Prof. Moncef Gabbouj, Dr. Esin Guldogan
Keywords: Content-based video analysis, machine learning, supervised classification, support vector machine, multimodality, multimodal fusion, genetic algorithms

Effective automatic analysis methods are required for the thorough utilization of extensive video collections. Content-based video analysis provides means for automated content interpretation. Popular approach for content-based video analysis is to consider it as a machine learning problem and apply supervised classification methods on low-level features extracted from the video key frames. Video recordings are usually accompanied by audio streams carrying complementary information, which can be a valuable asset in aiding the semantic analysis process. Fusion of information from visual and audio modalities have thus been studied in the literature. Multimodal fusion is a nontrivial task due to processing timing, synchronization, and modality applicability variance related issues. The recognition of the surrounding context from video recordings offers interesting possibilities for improved context awareness of mobile devices. For instance, the user profile of the mobile device can be adjusted, and tailored services provided according to the context.

In this thesis, a content-based video context recognition framework is presented. The framework utilizes visual and audio modalities. Support vector machine classifiers trained for six visual low-level features are used for recognition in the visual modality. The acoustic context recognition is provided by a recently introduced audio-based context recognition system. SVM-based fuser along with five rule-based fusion methods are integrated into the framework. In rule-based fusion, the separate classifiers are weighted with genetic algorithm optimized weights using presented weighting functions. The framework and the fusion approaches are evaluated on real-world video data recorded from 21 daily life contexts. Multimodal recognition is shown to outperform both unimodal approaches. The highest correct classification rate is achieved with SVM-based fusion.

PREFACE

This thesis work has been carried out in the MUVIS team at the Department of Signal Processing at Tampere University of Technology with the funding from TTY tukisäätiö. I would like to thank my supervisors Prof. Moncef Gabbouj and Dr. Esin Guldogan for their guidance along the thesis project. Miska Hannuksela and Antti Eronen at Nokia Research Center Tampere deserve special thanks for arranging the thesis project, as do Toni Heittola and Tuomas Virtanen of Audio Research Group (ARG) at the Department of Signal Processing for their assistance with audio-based context recognition. I would also like to express my gratitude for the members of the MUVIS team and ARG – especially the former ARG leader Anssi Klapuri.

I want to thank Johanna, my family, and friends for their love and support.

Finally, I would like to dedicate this work to the memory of my brother Jussi Roininen.

18.10.2010

Mikko Roininen

CONTENTS

ABBREVIATIONS AND NOTATION

ARM	association rule mining
CCA	canonical correlation analysis
CFA	cross-modal factor analysis
D-S	Dempster-Shafer
DBN	dynamic Bayesian network
DCT	discrete cosine transform
DT	decision tree
DTFT	discrete-time Fourier transform
EHD7	MPEG-7 edge histogram
EKF	extended Kalman filter
EM	expectation maximization
GA	genetic algorithm
GAlib	a C++ genetic algorithm library
GLCM	gray-level co-occurrence matrix
GMM	Gaussian mixture model
GPS	Global Positioning System
HHMM	hierarchical hidden Markov model
HMM	hidden Markov model
LDA	linear discriminant analysis
LIBSVM	a C++ SVM library
LPC	linear-predictive coding
LSA	latent semantic analysis
MFCC	mel-frequency cepstral coefficient
MR	manifold ranking
MUVIS	Multimedia Video Indexing and Retrieval System
NN	neural network
OpenCV	an open source computer vision library
ORDC	ordinal co-occurrence matrix
PCA	principle component analysis
QDA	quadratic discriminant analysis
ROI	region of interest
SMC	sequential Monte Carlo
SVM	support vector machine
SVD	singular vector decomposition
ZCR	zero-crossing rate
$\mathbf{x}, \mathbf{A}, ...$	boldface used for (column) vectors and matrices
ω	class

l	likelihood
$\check{\mathbf{l}}_{(j)}$	vector of likelihoods excluding the jth element of a vector of likelihoods $\mathbf{l}$
$\tilde{l}$	weighted likelihood
$P(\cdot)$	probability
$p(\cdot)$	probability density
$P(\mathbf{x}\|\boldsymbol{\theta})$	the conditional probability of $\mathbf{x}$ given $\boldsymbol{\theta}$
$p(\mathbf{x}\|\boldsymbol{\theta})$	the conditional probability density of $\mathbf{x}$ given $\boldsymbol{\theta}$
w	weight

1. INTRODUCTION

In recent years, the video medium has increasingly shifted from an one-way information transmission medium of broadcast companies to a personal expression and communication medium of everyday people. This is due to the introduction of highly popular social web services such as YouTube and Facebook, as well as the widespreadness of mobile multimedia devices and digital camcorders. According to statistics provided by the service about 2 billion videos are being watched and hundreds of thousands added to YouTube every day. This type of massive and ever-growing video databases require effective indexing and retrieval methods to be utilized to their full potential. The indexing has traditionally been performed manually, which nonetheless has its practical limits in terms of database size and growth pace. Traditional text-based search approaches are also limited to video databases with textual metadata or tags. Moreover, even if textual information is available, automatic metadata is limited in its descriptive power and user-created tags are unreliable, subjective, and need collective manual work to maintain.

Content-based semantic video analysis provides the means for automatic video indexing. As the analysis is based on the video content, no additional information or manual intervention is required. By utilizing features computed from the media content, considerably higher level of objectivity can be achieved compared to user-created tags. Combining information from multiple modalities such as video and audio allows more versatile and robust analysis of the semantic content. Different modalities can reveal diverse information from the data and support the conclusions drawn from other modalities. However, successful multimodal fusion is a task far from trivial. All the separate streams need to be processed in parallel, synchronized temporally, emphasized according to their applicability to the task, as well as combined in a sensible manner at the right point of processing.

Machine learning has been widely adopted for the classification of diverse information from individual data streams. It can be applied to multimodal fusion as well by combining the data or classifier decisions of the separate modalities, and using them as training data for a fusing classifier. Another popular option is to weight the modalities according to their applicability to a task, and use simple rules to achieve a combined decision among the weighted classifiers. In this case, reasonable weight values have to be chosen. Genetic algorithms (GA) offer automatic tools for

this type of an optimization task. GA-based search optimizes a set of parameters with respect to an evaluation task by applying the ideas of evolution and natural selection.

The combination of low-cost storage, decreasing size of video-capable mobile devices, and the development of content-based video analysis techniques has led to increased interest in the research of visual lifelogging. Visual lifelogging is the process of passively capturing video or images from everyday life. The unedited, continuously recorded data is stored in a personal collection, which can be used for retrieval of precious moments in life, as an aid for people with memory loss, or for safety and legal purposes. Understandably, the data amount of such a collection quickly becomes enormous and leads to the need of automatic content-based indexing.

An interesting application field for content-based semantic analysis of continuous video recordings from mobile devices is context awareness, which means adapting the properties of a mobile device according to the operation context. A simple example would be the automatic adjustment of a mobile phone usage profile according to the detected environment: In a crowded and noisy environment the phone could switch to a louder tone profile, whereas in more quiet contexts the profile would be switched back to normal. At specific recognized contexts, such as hospitals, aeroplanes, or theater performances, the phone could even switch itself off. More complex adaption forms include automatic offering of services based on the context, such as weather information, while outdoors, navigation system, while driving a car, or product and pricing comparison services, while shopping. Context can be estimated to some extend with positioning systems, such as GPS. However, positioning data cannot reveal the momentary situation in the estimated location, and additionally the current positioning systems are not applicable indoors. Audiovisual data is of dynamic nature and conveys rich instantaneous contextual information regardless of being indoors or outdoors.

This thesis presents a modular framework for multimodal content-based video context recognition. The framework uses visual features extracted from the key frames of continuous video recordings to train support vector machine (SVM) classifiers for distinguishing between different audiovisual contexts. Additionally, the framework fuses the recognition information of the internal classifiers with context likelihoods of an external audio-based context-recognition system. 5 rule-based fusion schemes and an SVM-based fuser are integrated into the system. The rule-based fusers are weighted with genetic algorithm optimized weights. Specific weighting functions for the rule-based fusers are presented.

The context recognition performance of the individual classifiers, the external audio-based system, as well as the different fusion approaches are evaluated on continuous audiovisual recordings from 21 everyday contexts. The contexts have

been chosen to represent environments and situations, where people would typically record video with a mobile multimedia device. Based on the best performing classifier combinations, optimal fixed weights are calculated for each rule-based fusion method.

The rest of the thesis is organized as follows. Chapter 2 presents the related work from literature and the theoretical background for the implementation. Chapter 3 describes the framework structure and implementation details. In chapter 4 the context recognition performance of the framework is evaluated with practical experimentations on real-world video data. Chapter 5 concludes the work and discusses possible directions for future improvements.

2. MULTIMODAL CONTENT-BASED SEMANTIC VIDEO ANALYSIS

Content-based semantic video analysis means automatic acquisition of high-level semantic information from videos using tools such as computer vision, signal processing, and machine learning. As the information is acquired directly from the video stream in an automatic manner, manual preparations or providing additional data is not necessary. The contents of supplementary parallel information streams such as audio and positioning information can also be studied to aid the analysis. This is known as multimodal content-based semantic video analysis.

One of the key advantages of content-based video analysis is that it allows automatic content-based indexing of video collections. Using this kind of indexing, videos can be retrieved much more effectively than is possible with traditional text retrieval methods. As the indexing is done automatically, it's applicable also to databases too vast and rapidly expanding for manual indexing.

2.1 Feature extraction

Content-based analysis of a multimedia database begins with feature extraction. Feature extraction is the process of deriving distinctive information in a compact form about an object or a set of data, e.g. the key frames of a video. Typically features are handled in a vector form, where each element represents some property. This property can be measured on a continuous scale, such as the average color wavelength of an image or the fundamental frequency of a sound frame, or on a discrete scale, such as the quantized intensity value of an image pixel or thresholded sound frame energy.

Feature extraction is a crucial part of a video analysis system. If the features extracted from the input data are of low discriminability in the first place, the system performance cannot be optimal no matter what post-processing methods are applied at later stages of processing. The extracted features can be categorized according to their abstraction level, scope of analysis, and modality as described in the following three sections.

2.1.1 Abstraction levels of features

In the multimedia analysis literature various abstraction level categorizations have been used for different tasks and modalities. The most common one is the simple distinction between low-level features and high-level features or semantic descriptors. Low-level features describe the objectively measurable properties of the media directly, whereas high-level features describe the meaning and purpose of the media contents [21]. However, a finer categorization granularity has been reported successful in some tasks [25, 47, 96].

Low-level features

Low-level or primitive features are commonly used for various multimedia analysis tasks. They are generally easy to extract and objective, but lack the sense of semantics of the multimedia content they describe - problem known as the *semantic gap* explained in more detail in section 2.2.7. As an example, according to a low-level color feature the images of a London Routemaster double-decker bus and a red rose might be highly similar although they usually have no semantic connection whatsoever.

Low-level features typically mimic the human perception for assessing the similarity or dissimilarity between data objects. The features can describe e.g. the motion in a video, color, texture, shape, or spatial location in an image or a video frame, as well as pitch, energy distribution, or zero crossing rate of audio data [21]. Simple statistical measures such as mean, variance, kurtosis, and skewness are used to represent the property distributions as real numbers.

High-level descriptors

High-level descriptors (also known as logical, derived, semantic descriptors/features) try to describe the content on semantic level. In contrast to low-level features they generally cannot be automatically extracted from the data, but require prior knowledge in some form [21]. Due to their nature high-level features are highly domain and task specific. The price of higher semantic interpretation capability is the decrease in objectivity and certainty.

High-level descriptors are usually an intelligent combination of low-level features (possibly between different modalities) and external knowledge embedded in an appropriate form. As they describe the semantics of the content directly and distinguish between different semantically interesting concept classes, high-level descriptors can be used on a per-concept basis. This leads to binary detection of the presence of each semantic concept in the multimedia item under analysis - a procedure commonly adopted in video concept detection frameworks in the recent years

[14, 17, 38, 75, 76, 88, 91, 97]. This approach also has the additional advantage of being able to use an optimized set of descriptors for each concept [6].

Mid-level representations

Different representations between the low-level features and the high-level semantic descriptors have been used in order to get a robust linkage to both the low-level features and the high-level concepts. This linkage can aid in mapping the low-level information to the semantic content.

In [25] sport video specific mid-level features such as camera motion patterns, action regions, and field shape properties are derived from a set of low-level visual features. Similarly, in [96] multimodal mid-level representations such as dialogue models are utilized in movie affective content analysis. Li and Tan [47] generate mid-level concepts such as video shot and face appearance from multimodal low-level features.

2.1.2 Feature analysis scope

Feature analysis can be done in various spatial and temporal scopes. With the proper choice of scope certain task-related aspects can be emphasized to make the analysis more suitable for the task.

Spatial scope

Spatial scope means the scale of analysis at one frame at a time instant. The scope can range from global features extracted from the whole analysis unit to highly local features considering only certain specific parts of the analysis unit. Choosing the right spatial scope for the features can highly boost the discrimination power of the features. As an example, if the features for a person identification task are extracted from a detected face region instead of the whole video frame, the features should be a lot more representative and robust to the environment.

For the visual modality spatial scope means choosing, whether the features are extracted globally from the whole frame, certain regions of interest (ROI) determined by some form of segmentation or object recognition, salient points, or fixed segments or points.

For the audio modality spatial scope can be regarded as the frequency domain scale and segmentation of the feature analysis. The analysis can be done for instance globally for the whole frequency band, for fixed or intelligently chosen subbands, or only to the harmonic parts of the sound after harmonic detection. Harmonic analysis can boost the performance in musical tasks and band-limited analysis can improve the robustness to wideband noise.

Temporal scope

Temporal scope of feature extraction means the granularity and the momentariness of the data analysis units with regard to time. Data capture rates can provide natural guidelines for the temporal granularity. However, in video analysis features are usually not extracted at each frame. This reduces the redundancy of the features and also decreases the computational constraints as features need to be processed less often. Common granularity choice for video content with a distinctive shot-based structure is to use intelligent shot segmentation and to extract the features on per segment basis. For non-structured and non-edited user created video material it's more suitable to extract the features from specific key frames of the video.

In the case of audio, a single sample wouldn't even provide enough information for the extraction of any practical acoustic features. By considering a weighted time window around the time instant, audio analysis becomes feasible. If the characteristics of a sound are assumed to stay constant within the time window, the frequency contents of the sound can be estimated using some frequency transform such as the discrete-time Fourier transform (DTFT). Longer windows offer higher frequency resolution, but reduce the time resolution. This can be countered to some extent by using overlapping of consecutive windows. Due to the higher capture rate and the lower information content of single samples of audio data, a short-time windowed audio frame is usually considered as the acoustic counterpart of a temporally static video frame. After all, in contrast to video the concept of an instantaneous sound lacking any time information makes really no sense.

Higher-level temporal relations of consecutive audio or video frames may be considered as well. This is the basic requirement for features describing temporal changes of some property such as position and speed of an object, or pitch or amplitude envelope of a sound.

2.1.3 Features of different modalities

Different modalities give different type of information about the captured content and context. The most commonly used modalities in semantic video analysis are visual, acoustic, and textual, all of which will be described in more detail later in this section.

Additional modalities - each having its own advantages and disadvantages - have been used in multimedia indexing and classification tasks. They include for example service point identification data (WLAN, Bluetooth)[39], light, temperature, infrared, and acceleration sensor data [24], time stamps, sensor location, geographical positioning data, and weather information [2], as well as orientation sensor data, wireless identification tags, and ultrasound [51].

Visual features

Visual data is rich in information, but also highly directional, and prone to occlusion, as only the content visible in the field of view of the camera is captured. It's also sensitive to varying lighting conditions, and can be cumbersome to interpret semantically depending on the content and task.

Visual features can describe the content in terms of low-level information such as color, texture, edges, shapes, motion, or more domain specific higher-level information including the presence and location of skin color, faces, sky region, or subregions containing text, as well as specific motion patterns of shapes, objects, or the camera.

Acoustic features

Audio data can also carry a lot of information, is sensed from every direction by nature, and doesn't depend on direct unobstructed path between the sensor and the source. Some of the main disadvantages of audio are vulnerability to noise, challenges of source detection and separation, as well as the possibility of information loss in certain multisound scenarios as quiet sounds get masked by louder ones. Especially speech can be highly informative, but needs manual labor or a recognition system to be utilized and is bound to a language.

Different acoustic features include low-level auditory features such as frame energy, spectrum centroid, spectrum flatness, zero-crossing rate (ZCR), harmonic spectrum centroid, various onset detection features, as well as mel-frequency cepstral coefficients (MFCC) and linear-predictive coding (LPC) coefficients along with their first and second order differentials [30]. Some of the higher abstraction level acoustic features include the presence of music or some distinctive sound event, voice activity, as well as prominent fundamental frequency and music tempo estimates.

Textual features

Textual data has the ability to contain easily interpreted information in a compact form. However, textual data requires manual annotations or recognition from other modalities. Textual information can be included in the form of scripts, subtitles, automatically added or user created tags and descriptions, various metadata from the capturing devices and post-processing phases, with the use of optical character recognition (OCR) in visual frames or automatic speech recognition (ASR) in the audio modality.

2.2 Semantic video concept recognition

Semantic video concept recognition is a popular semantic video analysis paradigm, which is based on the use of hierarchical detectors and their relations to detect different semantic concepts at various abstraction levels. A concept detector determines the probability of a high-level concept presence in a video segment by modeling the correspondence between low-level visual features and high-level semantic concepts using supervised machine learning [11].

Snoek and Worring [76] have defined the term *semantic concept* as "an objective linguistic description of an observable entity". Semantic concepts can thus range from generic object classes to particular members of a category such as specific persons and from single events and environments to complex chains of events. A combination or intersection of low-level concepts can also be regarded as a higher-level concept.

2.2.1 Applications of semantic video analysis

The automatic semantical analysis of video data has numerous application domains as it streamlines the information search in big video databases and thus enables new ways of utilizing huge collections of video data. In the literature several possible applications ranging from general to highly specific have been described for an intelligent video content analysis framework:

- *Assistive applications:* Passive continuous recording and storing of everyday life - so called *Lifelogging* - has gained increased research interest with the recording devices and storage space constantly becoming less expensive. This kind of data would rapidly grow into proportions impossible to handle manually, and is highly sparse in terms of interesting content. Automatic analysis would thus be the only option for summarizing the recordings and retrieving desired content. The usefulness of such a scenario is evident for the elderly and people with memory-related disabilities. The personal life-log could be used for recalling past events, for instance locating lost keys or ensuring that prescribed medicine was taken on time. [45, 51]. Lifelogging is discussed more thoroughly in section 2.6.

- *Automatic movie summarization:* Automatic semantic analysis would enable automating the process of movie summarization based on some criteria and constraints. A practical example of such summarization are movie trailers, which consist of excerpts invoking strong emotional reactions, but at the same time try not to give away too much information about the movie plot. [73]

- *Automatic sports video analysis:* The summarization and statistics creation of sports recordings as well as retrieval of specific events or persons (e.g. for performance analysis for training purposes) could be eased with an intelligent content-based semantic analysis framework. [76, 102]

- *Context awareness:* An online system for content-based video analysis could transform audiovisual sensory data into semantic information about the recording environment and adapt various properties of a mobile device (e.g. the ringtone and volume profile) according to the contextual information. [39]

- *Copyright infringement detection:* With automatic content-based video analysis it would be possible to search for copied material not only among video files in a database but also copied segments within longer video files. [76]

- *Managing personal collections:* With digital cameras and camcorders having surpassed their film counterparts, the growth of personal multimedia archives has sped up drastically as producing and storing new recorded content has become trivial and virtually free. As a consequence from this growth, searching for specific content has become more and more challenging and burdensome as one needs to go through bigger amounts of recorded material. An automatic semantic analysis framework would allow automated indexing of the data collection, and retrieval based on the indexing. Thus, no manual tagging or wading through the collection would be needed. [51]

- *Managing professional broadcast archives:* The media archives of national and commercial broadcast companies and the like are massive and expand constantly at a rapid pace. Managing these archives of both edited and raw video material is practically impossible without a versatile index of the contents. Manual annotation of the ever increasing collections of data is a tedious task. Automatic content-based indexing would provide a feasible solution for retrieval of documents from the archives. For such heterogenic collections of video data extremely generic detection of diverse semantic concepts is needed. [5, 76]

- *Social and collaborative sharing:* Social multimedia sharing services such as YouTube would benefit from automatic content analysis as it would provide tools for automatic tagging of content. This would greatly enhance the means of offering the users content similar to certain material or related to some high-level search criteria such as names of people or locations. [51]

- *Surveillance:* Continuously recorded surveillance videos produce extensive

amounts of content. Automatic tagging of significant segments (e.g. the presence of an unauthorized or suspicious person in a space or a traffic accident in case of a transportation video) is highly beneficial for efficient analysis of the surveillance material. [15]

2.2.2 Video context recognition

Content-based video context recognition can be seen as a subproblem of generic semantic concept recognition. In video context recognition the concepts are limited to a finite set of different types of audio-visual environments or location types around the recording device. The contexts can be defined as overlapping or mutually exclusive, the latter case being simpler to handle as the recognition output is always a single context.

Positioning methods such as Global Positioning System (GPS) can give valuable information about the location of a person, but they are not usable in certain situations (e.g. indoors in the case of GPS, although indoors positioning has also been studied for context estimation without visual data [57]) and only give geographical information. By this information alone it might not be possible to distinguish between being in the audience of a football game, track and field athletics competition, and a stadium rock concert.

In the literature various approaches related to visual and video context recognition have been proposed. Several video concept recognition systems address also concepts regarded as contexts. However, they are usually not restricted to mere contexts [11, 14, 38, 48] and additionally tend to be tuned only for the broadcast video domain and not for user created video content [17, 75, 88, 91, 97]. In [69] the authors match locations in feature-length films. Nevertheless, they only consider shot-based edited professional-quality video material and the approach concentrates on matching the same exact locations shown throughout the films instead of categorizing between different generic environments.

Blighe and O'Conner in [8] described a framework for recognizing real-world locations from passively captured images. The location classes consist of images from one particular location and not a general context. The framework uses the same sensor-equipped passive still image capturing device, Microsoft SenseCam, as the system studied in [40], where location information is provided with a GPS unit rather than estimated from the visual content. GPS along with other sensors is also used in [1] with content-based analysis only used for conversation detection. Blum et al. [9] classify between 8 mutually exclusive everyday environments in images taken once a minute using WiFi access point information and audio data.

2.2.3 Genericity of semantic video analysis

In recent years the focus of content-based video concept recognition has increasingly shifted from constrained approaches towards more generic systems. Generic systems aim at utilizing the least possible amount of domain knowledge about the video data and are thus less dependent on the video content, leading to wider applicability and easier extensibility - a natural direction for a gradually maturing research field.

Unfortunately, the price of generality is nonoptimal performance at specific tasks. Thus, domain knowledge still has its uses as the domain specific tasks are attempted to be solved in the most efficient and robust way. If the final application domain is fixed it makes no sense not to exploit the domain information and tune the system for the domain at hand. Domain knowledge can be incorporated by deriving domain specific mid-level representations or heuristic rules, which can improve the framework accuracy by either improving the feature representations or optimally filtering the data set [17].

Wickramaratna et al. [90] have used domain specific mid-level features (e.g. grass area ratio) to detect goal events in soccer games. In [20] American football games are analyzed according to camera view and play type. [100] uses face detection and tracking and script analysis to identify characters in feature-length films. Face detection is also used in [47], although they aim at a fairly generic system applicable to multiple domains. Smeaton et al. [73] use shot properties as well as visual and aural features effective at detecting exciting sequences of a movie in order to automatically form action movie trailers.

2.2.4 Temporal dependencies

Temporal dependencies or relations can be utilized to refine multimodal semantic concept detection results by analyzing, how the concepts behave in consecutive analysis units as concepts and especially events usually span over multiple shots or key frames [16]. With the aid of temporal consideration video frame regions and points can be tracked, camera motion estimated, pixels grouped according to motion speed and direction, and activity in general detected and measured [76]. With acoustic data temporal relations can be used for source tracking as well - especially with multiple sensors. Some tasks such as automatic speech recognition wouldn't even make sense without considering the temporal progression of the classification. Temporal dependencies can also be used to filter out outlier misclassifications.

HMMs have been extensively applied for inherently considering temporal relations during the classification [19, 27, 36, 59, 62, 94]. In [59] a framework is proposed that jointly fuses HMM-classified unimodal streams and exploits temporal dependencies using a Bayesian network to detect semantic concepts in the domain of news broad-

cast videos. Chen et al. [15] perform temporal pattern analysis as the second step of a three-stage video event detection framework. The aim of the step is to identify significant temporal patterns to be used as temporal features for data mining, and to perform data reduction. Jiang et al. [38] temporally track region-based visual features jointly with background audio features to recognize semantic concepts from video. In the proposal by Weng and Chang [88] temporal relations are explored simultaneously with inter-contextual relation in a scalable manner. The proposed framework is independent of the individual classifier types. In [80] a semi-supervised video indexing framework using filtering-based temporal consistency exploiting is proposed. The framework uses decision voting for fusion. Liu et al. [50] proposed a post-filtering framework for association and temporal analysis for semantic concept detection.

2.2.5 Concept relations

Concept relations represent the semantic linkage between co-occurring multimodal features and concepts. Modeling the relations can be done explicitly e.g. using directed or undirected graphs, or implicitly e.g. by machine learning, manual adjustments, and greedy partitioning. To some extent the relations can also be mined and new combinatory concepts derived from external ontologies. Relations modeling supports inference: the likelihood of a detected concept can be adjusted according to detected co-occurring concepts. [76]. For instance, the simultaneous detection of concepts like *beach*, *forest*, *lake*, and *birds* could increase the likelihood of *outdoors* and decrease the likelihood of *office*.

Wu et al. [91] proposed a multimodal semantic video concept detection framework based on 3rd order tensors, which intuitively express the concept relations. In [48] the authors proposed a framework for semantic video concept detection based on *association rule mining* (ARM), an accurate and efficient concept relation and temporal dependencies mining method. In their experiments with the detection of 15 everyday concepts the ARM-based approach outperformed NNs, SVMs, and DTs. As mentioned in section 2.2.4 the framework by Weng and Chuang [88] utilizes both temporal and inter-concept relations, as does the framework in [50]. The semantic video analysis paradigm proposed by Qi et al. [62] simultaneously models the individual concepts and their correlations, which according to the authors prevents error propagation between the concept detection and relation analysis. Yanagawa et al. [97] use in their video concept detection framework a graph-based method to learn dependencies between 374 semantic concepts.

2.2.6 Evaluation metrics

Various evaluation metrics have been developed for measuring and comparing the results of classification and retrieval tasks in general. One of the simplest measures is the *correct classification rate*. It is calculated simply as the percentage of correctly classified samples in an unseen test dataset. Correct classification rate presents a quick estimate on the overall performance of the classification system, but doesn't give any information about the individual classes. A related metric for binary problems is the *classification accuracy*. It is defined as

$$classification\ acc. = \frac{identified + rejected}{identified + misidentified + missed + rejected}, \tag{2.1}$$

where $identified$ corresponds to true positives, $rejected$ to true negatives, $missed$ to false negatives, and $misidentified$ to false positives.

Precision and *recall* are two widely used metrics for classification and information retrieval [15]. Precision describes, how well the items actually belonging to a certain class are identified, and is defined as

$$precision = \frac{identified}{identified + misidentified}. \tag{2.2}$$

Precision of a class only considers the test items labeled by the system as belonging to the class. Hence, high precision values can be achieved with a strict screening process, where only the most clear cases are marked as belonging to the examined class. Recall, on the other hand, describes how well the instances of the examined class are discovered from the test data. Recall is defined as

$$recall = \frac{identified}{identified + missed}. \tag{2.3}$$

Recall of a class can be improved by having the system label the test items as belonging to the examined class from the slightest hint so as many as possible items of the examined class are found.

As both the precision and recall depend on the class acceptance and thus from each other they are often given as a precision–recall graph [37]. It should be noted that in practical mutually exclusive classification tasks it's not possible to improve the overall precision or recall much by class-wise acceptance tuning, as accepting more samples to be classified to one class means rejecting them from other classes. Precision and recall have been criticized for their dependency on the relevant class size, class amount, and test set size [37].

F-measure combines precision and recall into one measure [44]. It can be formal-

ized as

$$F_\alpha = \frac{1}{\alpha \cdot \frac{1}{recall} + (1-\alpha) \cdot \frac{1}{precision}}. \tag{2.4}$$

$F_{0.5}$ would weight the two metrics equally. Larger values of α give more weight to recall and smaller values emphasize precision.

Several other evaluation metrics exist for specific multimedia analysis tasks. These include the *NIST average precision* and *mean average precision* metrics for evaluating the performance of a retrieval task from an ordered list of retrieved items, *mean distance from track*, *detection rate*, and *false positive rate* for tracking related tasks, *certainty*, *accuracy*, and *timeliness* for information fusion, as well as *false acceptance rate* and *false rejection rate* for biometric verification. [2]

2.2.7 Challenges in semantic video analysis

Replicating the visual and aural stimuli based abstraction and reasoning process within the human brain is a task far from trivial. The paradigms used for unimodal or multimodal concept recognition suffer inherently from certain problems, some of which have already been mostly solved - some still remaining an open question.

The sensory gap

The problem of within-concept feature variance due to different sensing conditions (e.g. lighting, viewpoint, visual background, varying room acoustics, background noises) is called the sensory gap [76]. The effects of the sensory gap can be reduced by using features invariant to the conditions. Nonetheless, at some point the amount of this invariance begins to affect the discriminatory power of the feature. Thus a balance needs to be found based on the features and the application.

Various visual and acoustic features with different levels of condition invariance have been developed. Due to the matureness of the feature extraction field the main research focus has shifted from bridging the sensory gap to that of the so called semantic gap.

The semantic gap

The semantic gap is generally acknowledged as the most fundamental challenge of general semantic video analysis [12, 17, 20, 46, 72, 76, 87, 91]. Hence its causes and possible solutions have been thoroughly discussed in the literature. Snoek and Worring [76] define it as "The lack of correspondence between the low-level features that machines extract from video and the high-level conceptual interpretations a human gives to the data in a given situation." In the work by Chen et al. [17] the main research fields for addressing the semantic gap are categorized as follows.

- *Syntactic analysis*, i.e. search for optimal partitioning of the video data and improved features harnessing data from one or multiple modalities.
- *Decision-making process*, i.e. optimizing the mapping of feature and related data to the semantic concepts in the system output with techniques such as machine learning and temporal relation analysis.
- *Domain-related modeling*, i.e. using diverse domain specific information to boost the detection performance for a limited set of tasks.

Wu et al. [91] believe that the combination of fusing information from multiple modalities and taking semantic context into account is likely to bridge the semantic gap. Snoek and Worring believe that the answer to the problem lies in automated detection, selection, and interactive usage of semantic concepts [76]. Despite massive research effort the issue still remains unsolved in the general case.

Data imbalance

For the training of several classification algorithms, performance-wise the ideal case is when the amount of positive samples (corresponding to the requested class) roughly equals the amount of negative ones (corresponding to any but the requested class). The negative effects of data imbalance can be reduced for example by using classifiers robust to the problem or by adopting different data sampling methods to reduce the size difference of the negative and positive sample sets. Two common cases of data imbalance are *positive-to-negative samples ratio* and *class imbalance*.

With the increased amount of classes the relative size of the sample set corresponding to any one class decreases. This leads to increased difference between the amount of positive and negative samples. Chen et al. [16] use a bootstrapped sampling scheme presented in their earlier work [90], where the negative sample pool is divided into subsets of roughly the size of the positive pool and multiple neural networks are trained using the positive set and each of the negative sets in turn. In [17] an eigenspace analysis utilizing subspace-based data pruning method is applied to tackle the problem.

Class imbalance typically encountered in rare event detection tasks is a form of data imbalance caused by the infrequency of interesting concepts in the data. A good example case is the detection of goal events from soccer game material - the class of goal event samples is typically extremely small compared to the rest of the data. Shyu et al. proposed the integration of distance- and rule-based data mining techniques in order to address the class imbalance problem [72].

Overfitting

Overfitting occurs as a classifier shows high performance classifying the training dataset, but fails to generalize well for unseen data. This indicates that the model parameters have been over-optimized and the classifier has learned irrelevant details from the training dataset, but poorly acquires the properties of the underlying probability distribution. To avoid overfitting the used learning methods shouldn't be overly complex and the training data should be representative enough.

Feature dimensionality problems

High-dimensional feature vectors can cause various problems in classification. One common problem is the so called *curse of dimensionality*, which means the exponential growth in the training data need with the increased feature vector length for retaining proportional accuracy in the feature space sampling density. As gathering huge amounts of training data can be costly and too drastic dimensionality limitation can also lead to poor performance, an appropriate balance should be found for the ratio of the data amount and dimensionality.

Along with the curse of dimensionality another major problem with high feature vector dimensionality is that the computational complexity of classification algorithms is typically at least quadratically complex with respect to the number of features. Hence, feature filtering and dimensionality reduction could potentially speed up the processing drastically [29]. Shorter features also require less storage [72]. Various dimensionality reduction techniques have been developed to alleviate the problems with high-dimensional feature vectors.

Atrey et al. describe dimensionality reduction methods commonly used in multimodal information fusion in multimedia analysis tasks. *Principle component analysis* (PCA) projects high-dimensional data into a lower-dimensional space by choosing the dimensions that minimize the squared error in reconstructing the original data. It has been reported not to work well with very high dimensionality. *Singular vector decomposition* (SVD) works by determining the eigenvectors that most represent the input feature set. It's an unsupervised method having no problems with high dimensionality. *Linear discriminant analysis* (LDA) is a supervised method for determining the optimal linear combination of features. It not only reduces the dimensionality, but can also be directly used for classification. In addition to these, particularly for dimensionality reduction developed techniques, some data relation analysis methods such as *Latent semantic analysis* (LSA) and *cross-modal factor analysis* (CFA) also offer dimensionality reduction capabilities. [2]

In [71] some additional dimensionality reduction methods are described. These include the *discrete cosine transform* (DCT) broadly used in image compression,

which approximates the data vector as a sum of cosine functions, and *quadratic discriminant analysis* (QDA), a quadratic extension of LDA. According to the authors both LDA and QDA are known to outperform PCA in classification tasks. They also mention *canonical correlation analysis* (CCA), a statistical approach combining linear dimensionality reduction and information fusion by computing maximally correlated linear projections. Shyu et al. [72] have recently proposed a semantic video event/concept detection framework that automatically reconstructs and refines the feature dimensionality. There refinement is based on the first dimension of typical negative eigenspace presenting the most data information. They report the automatic system typically reducing the dimensionality to 50 % of the original in practical experiments.

Normalization

Different feature extraction methods produce data with different dynamic ranges. This can lead to undesired emphasizing of certain features as for instance the widely used euclidean distance similarity measure is sensitive to magnitude. Different normalization methods exist against this phenomenon. A common convention is to shift and scale all the features to some fixed range - usually between 0 and 1 - based on the minimum and maximum values. Other options for the scaling are to use the mean and variance, or mutual ordering of the feature values. [76]

Missing or erroneous data

With big real-world databases it's practically inevitable to avoid noise and mislabeling in the data. In some cases some features might also be entirely missing. If these problems are too prominent, it can lead to deterioration of the system performance. In such a case the data should be filtered and only the good data used to build the classifier models. Gabrys and Ruta report two categories of this filtering: so called *data editing*, where the representative samples are chosen directly, and segmentation of the input space. [29]

2.3 Statistical classification

Statistical classification is a common term for supervised machine learning procedures, where a class label is assigned for an input data sample based on various compact and representative properties calculated from the sample. Classification involves a training phase where the classification system forms discriminative statistical class models from training data examples with known class labels.

Statistical classification is closely related to statistical regression and clustering. In all of these paradigms an optimal response is sought for an input sample. In the

case of regression the response is continuous and not a discrete class. Clustering is a form of unsupervised machine learning, where the system is not trained beforehand with labeled samples but has to find the classes from the unlabeled input data and assign the samples into these class clusters.

Statistical classifiers are widely adopted in content-based semantic multimedia analysis. This is due to the inherent classification nature of the tasks, but also as the analysis aims at extracting information from high semantic abstraction levels, which would be much more wearisome with mere clustering without any training examples. Additionally, as several analysis tasks aim at accomplishing functions trivial for a human, training data annotation is usually quite straightforward - although can be laborious with big databases.

2.3.1 Machine learning methods

Statistical classification is a broad research field with a huge amount of applications. Over the years numerous classification algorithm variations have been developed, each one focusing on some specific set of problems. Classification algorithms can be grouped into binary and multi-class algorithms. Binary classification is a much simpler and more extensively studied problem and thus contains a wider range of algorithms. However, a great deal of real-world problems require the use of more than two classes, thus encouraging the research of multi-class classification. Due to the profound work with binary classification several multiclass algorithms are in fact merely combining binary classifiers, but purely multi-class algorithms also exist. The rest of the chapter presents some of the most popular machine learning schemes used in video and multimedia classification.

Bayesian inference

Bayesian inference is a classical approach to the pattern classification problems. It estimates the most likely output classes based on probabilities of the data given an observation (a test sample) as stated in the *Bayes' formula*:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}. \tag{2.5}$$

Term $p(x|\omega_j)$ is the *state-conditional probability density* or *likelihood* of class ω_j producing an observation x, $P(\omega_j)$ the *prior probability* of each class, and $p(x)$ so called *evidence*, which ensures $P(\omega_j|x)$ being a true probability. The evidence is independent of the class and can thus be ignored in determining the most likely output class. $P(\omega_j|x)$ is called the *posterior probability*, the probability of the output class being ω_j given a feature x. The class ω_j maximizing the posterior probability is chosen as the system output. The choices can further be weighted by assigning

different costs for misclassification of different classes. [26]

The Bayesian inference approach has various advantages: Based on new observations, it can incrementally compute the probability of the hypothesis being true. The new observation is used to update the prior probability. With absence of empirical data, subjective probability estimates can also be used as the priors. Still, the success of the approach is highly dependent on the reliability of the likelihood estimates. [2]. The approach is the basis for several advanced algorithms and schemes, but it has also been applied directly, for instance to modality fusion in event detection in team sports videos [95] and in audio-visual speech recognition [60].

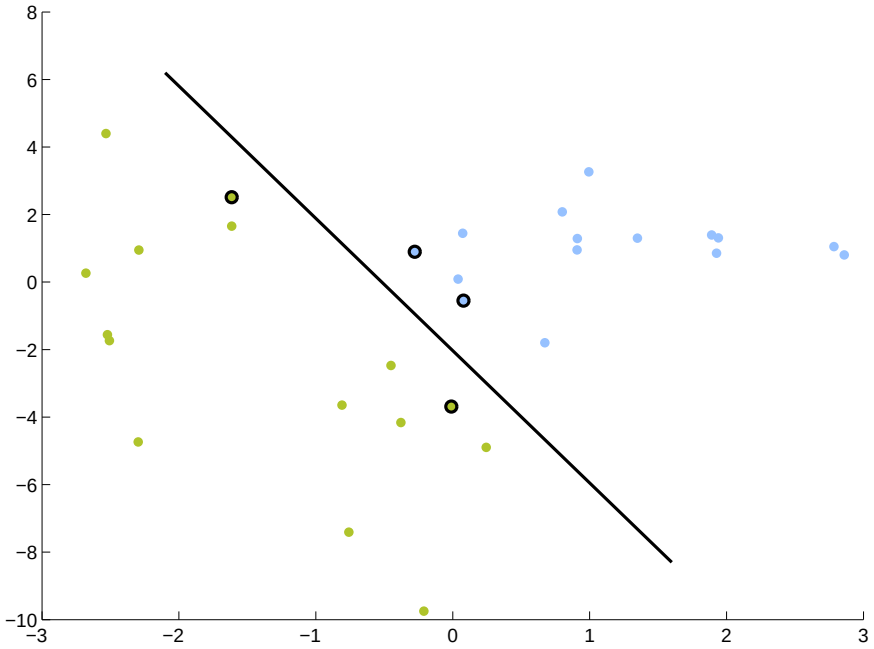

Figure 2.1: SVM decision boundary in a simple linearly separable binary classification problem. The support vectors of each class are encircled with a black edge.

Support vector machines

Support vector machines (SVM) [84] is a commonly used supervised learning method for classification. In SVM classification the class separation hyperplanes are placed by training to maximize the margin between the hyperplane and the closest input data vectors (the so called support vectors). SVMs can use kernel function transforms to map the data into a higher-dimensional space for more straightforward separation. In case the classes overlap and are nonseparable even after the kernel transform, cost functions are utilized to give higher penalty to features residing further away in the wrong side of the separating plane. Figure 2.1 shows a simple

linearly separable SVM hyperplane in a binary classification task.

SVMs are known for their satisfactory generalization performance and often higher classification accuracy compared to other well-known pattern recognition techniques such as *maximum likelihood* and *multilayer perceptron neural networks* [56], and *Gaussian mixture model* (GMM) and *manifold ranking* (MR) [98]. Moreover, they have been reported to work well with small training data amounts [15]. Additionally, according to [71] SVMs shouldn't, in theory, require an explicit dimensionality reduction step as they use the internal kernel transforms.

SVMs have been criticized for their decreased performance with increased training data sizes [72] and feature vector dimensions [55], as well as having low accuracy in tasks with high class-imbalance [92, 97] as the rare samples don't represent the true class distributions well enough. Important factors in the performance of an SVM classifier are the choice of the kernel function and its parameters. The features should also be roughly in the same dynamic range. [97]

In the field of multimedia categorization, the SVMs have been utilized in various applications such as multimedia semantic concept detection [74, 79, 97], interactive video retrieval [74, 77], video retrieval result reranking [35], web video categorization [98], automatic shot selection for action movie trailers [73], video concept co-occurrence relation modeling [3], visual lifelog everyday concept detection [11], and automatic semantic video annotation [80]

In [91], the idea of SVM has been extended from vectors to higher order tensors to form so called transductive support tensor machines. The algorithm trains classifiers efficiently utilizing semi-supervised learning techniques. Zhang et al. in [99] proposed an incremental SVM with fixed number of support vectors for large scale incremental learning in the scope of web video categorization.

Neural networks

Neural networks (NN) (also known as artificial neural networks to distinguish them from their biological counterparts) have been widely applied to diverse classification and event prediction problems due to their strength of identifying the relationship between predictor variables (inputs) and predicted variables (outputs) even when the relationship is far too complex to model with other mathematical approaches such as correlation. According to Ranawana and Palade [65] NNs have also proven to be the most popular intelligent multi-classifier combiner. NNs have been reported to have low accuracy in rare event detection due to slow convergence. [90]

Haykin in [32] has defined a neural network as follows: "A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network from its environment through a learning process.

2. Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge."

Neural networks consist of three types of layers of nodes or neurons and their connections. The first type of processing layer is the *input layer*, the node count of which corresponds to the amount of input variables. The last layer, the so called *output layer*, contains as many neurons as there are desired outputs (e.g. the amount of different classes in case of a classification task). The basic NN scheme is shown in Figure 2.2. All the other layers in between are called *hidden layers*. Nonlinear input-output mappings of various complexity can be approximated to any degree of accuracy depending on the inner network structure. This is done by iteratively adjusting the synaptic weights inside the network from a training data set to minimize the difference between the system output and training data labels [68]. According to Atrey et al. in [2] the network architecture at the hidden layers is an important factor for the success or failure of NN classification. They also claim that even though neural networks are suitable for high-dimensional problems and generate high-order nonlinear mapping, the proper choice of appropriate network architecture for a particular application is often demanding and training of a NN is relatively slow.

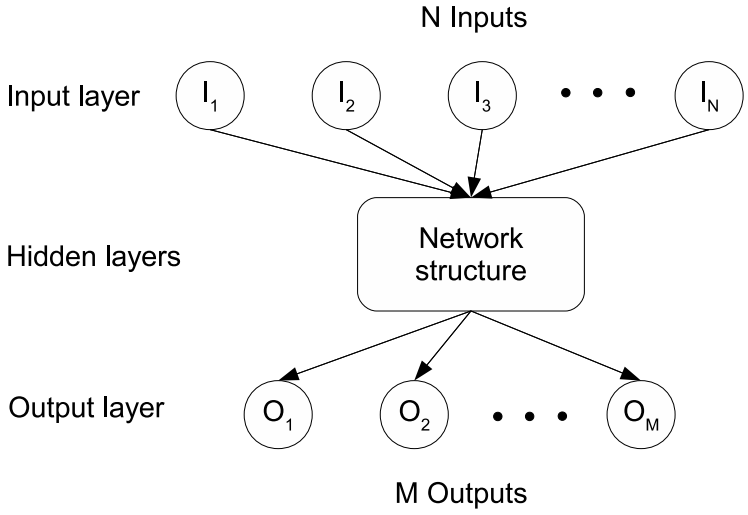

Figure 2.2: The basic neural network scheme [68].

Wickramaratna et al. in [90], have proposed a NN-based framework for soccer video goal event detection. They use a bootstrapped ensemble of neural networks to alleviate the class imbalance issue. Benmokhtar and Huet have used a extended version of NN to fuse the outputs of different concept detectors [6].

Hidden Markov Models

Hidden Markov models (HMM) are a popular modeling method for problems with inherent temporality. HMMs are a special case of dynamic Bayesian networks (DBN), which extend Bayesian inferencing to graphs. With this model each temporal system state depends only on the state at the previous time instant. HMM consist of nodes representing hidden states, connected by links with transition probabilities for moving between states, and having probabilities for emitting different visible states. After training the transition probabilities from training sequences using the *Forward-backward* or *Baum-Welch* algorithm, the hidden state sequence that most likely has produced the observed visible state sequence is chosen as the system output. The search of the most likely sequence is done with the *Viterbi* algorithm. [26]

Due to the temporal nature of recorded multimedia sequences, HMMs have been widely adopted for multimedia analysis. In traditional unimodal tasks especially the speech recognition community prefers to use HMMs [71]. Papadopoulos et al. [59] use HMMs to initial shot classification of unimodal streams in semantic video analysis. In [36] HMMs are used for analyzing the accelerometer sensor data of a lifelogging system in order to estimate user motion patterns. Ebadollahi et al. [27] have adopted a dynamic multi-concept approach for detecting semantic events in videos. They use HMMs to model the temporal changes in concept presence estimates. In [19] HMMs are utilized to classify broadcast video genre based on text and face detection. Qi et al. [62] recently proposed a new semantic video analysis paradigm using HMMs for exploiting the rich temporal information in videos. In [94] it's shown that the *hierarchical hidden Markov model* (HHMM) outperforms *frequent itemset mining*, *k-means clustering*, and traditional HMM in temporal pattern mining in large-scale video concept streams.

Decision trees

Decision trees (DT) are a computationally efficient group of algorithms capable of classifying data with a recursive process of simple attribute comparisons. The classification starts at a *root node*, from which *links* or *branches* lead to child nodes of the root. Appropriate branch is chosen based on some attribute or thresholded property of the data and the analysis moves to the corresponding child node. The child nodes are again linked to further nodes and the process continues recursively until a terminal or *leaf* node is reached. The leaf nodes represent different classes. The most popular DT algorithm for classification is the C4.5. [26]. In addition to low computational costs, DTs have the advantage of being able to select the representative feature items automatically [15].

Shyu et al. [72] have used the C4.5 trees for detecting soccer goal and corner kick events in soccer videos, and semantic concepts in news videos. In [52] a novel decision tree algorithm is proposed for image semantic analysis.

Heuristic rules

Classification with heuristic rules means using domain expertise and trial and error approach to come up with class decision rules, which are not strictly based on statistical classification or other formalized theories. Heuristic rules can eventually lead to a new classification scheme if their behavior can be predicted and formalized in the general case.

As stated in section 2.2.3 heuristic rules can be utilized to incorporate domain knowledge in domain specific problems. In [64] a heuristic consensus learning algorithm is proposed for clustering and classification of user-generated video and related metadata in social multimedia sharing services. Xu and Chua in [95] applied domain specific rules along with generic feature analysis for team sports video event detection. In [4], custom rules are used for semantic sports video indexing.

2.3.2 Semi-supervised learning

Various semi-supervised learning methods have been proposed more recently in addition to the massive amount of supervised and unsupervised machine learning techniques. These procedures are motivated by the idea of combining the advantages of supervised and unsupervised learning, specifically using a large training data set in a supervised-like fashion without the need to hand label large amounts of data. In other words, unlabeled data obeying certain assumptions is automatically exploited to help estimate the data distribution in order to improve learning performance [2]. With semi-supervised learning methods higher accuracy can be gained with less human effort compared to traditional classification methods. Choosing of appropriate models, features, kernels, and similarity functions is nevertheless needed to match the assumptions made for the unlabeled data [101].

Most widely used methods according to the survey in [101] include *expectation maximization* (EM) with *generative mixture models*, *self-training*, *co-training*, *transductive support vector machines*, and graph-based methods. Semi-supervised ideas have been used to improve traditional supervised techniques such as SVMs [7, 34, 63]. Other unlabeled data handling techniques related to semi-supervised learning are *transductive learning* , in which the unlabeled data is taken from the test data set, and *active learning* , where the learning algorithm selects unlabeled samples and asks the user for labels [2]. Semi-supervised learning has been successfully applied for example to semantic video analysis [28, 58, 80] and image retrieval [34].

2.4 Multimodal fusion

Multimodal fusion is a special case of combining information from multiple information streams. Modality fusion methods can be categorized according to the different types of information available for fusion at different stages of the multimodal analysis task. Shivappa et. al [71] present a 5-stage categorization with the following categories:

1. Signal enhancement and sensor level fusion techniques
2. Feature level fusion techniques
3. Classifier level fusion techniques
4. Decision level fusion techniques
5. Semantic level fusion techniques

Feature level fusion and decision level fusion have been most commonly utilized in the semantic video concept recognition literature [11, 38, 76, 78, 97]. The former method combines feature vectors before classification and the latter decisions of the individual classifiers. Signal enhancement and sensor level fusion strategies consist of methods such as noise suppression and beamforming with multiple sensors. Classifier level fusion fuses separately processed features inside a composite classifier. It differs from feature and decision fusion in that it doesn't work directly on features and the fusion happens before decision making. This type of fusion is only feasible with certain types of classifiers such as HMMs. In semantic level fusion, the fusion is done by combining the semantic interpretations. Semantic fusion has not been studied extensively due to the difficulties of processing information at such a high abstraction level. In addition, various hybrid fusion techniques exist combining fusion methods from multiple categories [2, 65].

2.4.1 Feature fusion

Feature fusion or so called early fusion collectively utilizes the information from different feature streams by combining the features and training a common classifier instead of training separate classifiers for each feature. Usually the integration is carried out by simply concatenating all feature vectors from different streams. Additionally, other methods such as summing the features have been used in the literature [38]. The advantage of this approach is the utilization of the correlation between different features at an early stage which helps in better task accomplishment [2]. Moreover, only one classifier needs to be trained [78]. Snoek et al. in [78]

have defined early fusion in the domain of multimodal video concept detection as: *Fusion scheme that integrates unimodal features before learning concepts.*

The main disadvantage is the curse of dimensionality due to high-dimensional feature vector combinations. Combining the features can also be problematic due to various representation differences [78]. The features might have completely different dynamic ranges and distributions, which may lead to erroneously favoring some features over some others [76]. To overcome this, normalization needs to be applied to the features before the combining process. Additionally, the capture properties might differ in terms of rate, phase offset, and rate variation in multimodal data streams, which leads to synchronization requirements before combining the vectors [93].

According to Tseng et al. in [82], early fusion is suitable in situations, where large amounts of training data are available and the samples are correlated and dependent. On the other hand, feature independence is required both in unimodal and multimodal cases as stated by Snoek and Worring in [76].

2.4.2 Decision fusion

Decision fusion or late fusion combines the different information streams in the semantic space after classifying the features separately. In other words, the fusion is performed on the outputs of the individual classifiers. In [78], the late fusion for multimodal video concept detection is defined as: *Fusion scheme that first reduces unimodal features to separately learned concept scores, then these scores are integrated to learn concepts.*

The drawback compared to early fusion is that each separate information stream needs its own learning stage and the fuser might need to be trained as well depending on the fusion method. Training the individual classifiers can nevertheless be faster than the learning process of one high-dimensional concatenated feature vector. This depends mostly on the scalability of the chosen classification procedure. According to Snoek et al. in [78], late fusion outperforms early fusion in semantic video concept detection with most of the concepts. On the other hand, the performance difference is more notable with the few concepts, where early fusion performs better.

Decision fusion methods

Different late fusion strategies can be utilized depending on the types of outputs of the classifiers. In [68], classifier outputs are divided into three categories: crisp class labels, class rankings, and soft scores.

Crisp class labels mean that only the label of the most likely class according to the classifier is given as output. Class ranking output consists also of labels, but

instead of only the most likely one, all the labels are ranked according to their likelihood of occurrence. In the case of soft score output, the likelihood score of each class is directly available. Soft scores offer the most information and it's easy to convert them into class rankings and further into crisp label outputs. It's much more difficult to travel into the opposite direction.

2.4.3 Hybrid fusion

Hybrid fusion is a combination of different fusion methods - typically early and late fusion in semantic video concept recognition tasks. Hybrid combination techniques try to utilize the advantages of the component fusion methods [2]. The combination strategies can range from simple sequential setups to complex hierarchies depending on the application needs. The main disadvantage is the increase in computational costs with more complex systems.

2.4.4 Fusion methods

Apart from categorizing fusion strategies by applying point and the type of information, the methods can be further categorized based on partitioning of the problem space into rule-based methods, classification-based methods, and estimation-based methods. The former two can be used for decision making based on an observation, and the last one for parameter or state estimation. [2]

Rule-based fusion

Rule-based fusion methods are computationally inexpensive and work well if the temporal alignment between different modalities is accurate. *Linear weighted fusion* is one of the most widely used methods due to its simplicity and applicability to both feature and decision level fusion. The general idea of linear weighted fusion can be formalized by using the sum operator as shown in

$$I = \sum_{i=1}^{n} w_i \times I_i, \tag{2.6}$$

or with the product operator as shown in

$$I = \prod_{i=1}^{n} I_i^{w_i}, \tag{2.7}$$

where I_i are the feature vectors of n data streams or decisions of the n classifiers, w_i the corresponding weights, and I the fusion output. The weight calculation and adjusting is the most decisive part for the success of the fusion. *Majority voting rule* is a special case of weighted combination with uniform weights. The fused

decision is the class, for which the majority of the classifiers make an equal decision. [2]

Minimum and *maximum* operators are also widely used for fusion due to their simplicity [65]. Among all the classifiers the maximum likelihood for each class can be searched, and the class corresponding to the maximum of maximums taken as the system output. This is analogous to trusting the most sure class estimation by any of the classifiers. Similarly, the minimum likelihoods of each class among the classifiers can be searched and the class corresponding to the maximum of these minimums taken as output. Thus the the least doubted class is chosen.

Similar to heuristic rules in classification, *custom-defined rules* are valuable in fusion cases, where the application domain is known. Custom rules can be added based on domain knowledge and requirements.

Classification-based fusion

In classification-based fusion the combination of separate information streams is carried out by statistical classification methods such as the ones presented in section 2.3.1. *Dempster-Shafer* (D-S) theory and *maximum entropy model* are two additional classification methods popularly used for fusion. Dempster-Shafer theory uses the concepts of *belief* and *plausibility* describing the evidence and uncertainty of a decision hypothesis, respectively. D-S theory is a generalization of the Bayesian inference procedure. It has been defined to handle better situations, where the classes are not mutually exclusive. The main drawback with the method are the handling difficulties of large number of combinations of hypotheses. Maximum entropy model is a statistical classifier based on information theory. The probability of an observation belonging to a class is based on the information content of the observation. [2]

Estimation-based fusion

In general, estimation-based fusion in multimedia analysis is utilized to better estimate the state of a moving object based on multimodal data. One typical example would be the tracking of a talking person based on the combination of face detection from video and sound-of-arrival estimates from speech.

Kalman filter is a popular and well-established fusion method applicable for real-time source localization and tracking by processing of dynamic low-level features. Each state estimate given by the method only depends on the previous state. The disadvantage of the method is that it's not suited for nonlinear problems. This restriction has been tackled in the so called *extended Kalman filter* (EKF) method also successfully applied to source localization for several years [66].

Another well-known set of estimation-based fusion methods are the *particle filters* or the *sequential Monte Carlo* (SMC) methods. In these methods a set of particles representing the random samples of the state variable propagate in the state-space and get weighted according to the latest sensory information. With sufficiently large number of particles the methods can be used to estimate the state distribution in the nonlinear and non-Gaussian state-space model. [2]

2.4.5 Diversity between modalities

The so called *diversity* between the different information streams has a significant impact on the performance of decision fusion [43, 65, 67]. Diversity is the degree of independence between the information streams to be fused. Diverse classifiers distinguish different regions of the input space and thus complement each other. One of the advantages of multimodal approaches is that the diversity between modalities tends to be higher than between classifiers of the same modality.

Ranawana and Palade in [65], described means of incorporating diversity into a fusion system. One option is to train multiple classifiers for subsamples of the input data. The output classes can also be grouped and different classifiers trained to distinguish between the classes within one group. With some training algorithms it's possible to incorporate randomness in terms of random initial weights. The training algorithm parameters, such as the number of neurons and hidden layers of a neural network, can also be varied. The authors also presented pairwise and non-pairwise diversity measures, but stated that neither individual classifier performances nor the diversity measures provided an foolproof indicator on the fused classifier performance. Rokach in [67] recently described additional diversification procedures such as partitioning of the search space and hybrid use of different methods in different subspaces.

2.4.6 Correlation between modalities

Even though the information streams to be fused are desired to be highly diverse, their correlation is also an important factor for the performance of the overall system. Specifically, the information streams having high confidence at a input space region are assumed correlated in this region for not causing conflicts at the fusion.

Atrey et al. [2] have reported methods for exploiting feature and decision level correlation. Feature correlation can be calculated using *correlation coefficient* and *mutual information*, in addition to LSA, CFA, and CCA mentioned in section 2.2.7. Correlation coefficient measures the strength and direction of a linear relationship between two features assumed to be independent and jointly follow the Gaussian distribution. Mutual information is an information theoretic measure representing

the amount of information one normally distributed feature reveals about another.

Decision correlation can be estimated in the form of *causal link analysis*, *causal strength*, and *agreement coefficient*. Causal link analysis iteratively finds likelihoods for the causal links of pairs of semantic events and their relative times of occurrence. Events often occurring close to each other get higher correlation likelihoods. Causal strength estimates the causes of correlation between two variables instead of measuring the correlation directly. Agreement coefficient measures, how concurring or contradictory is the evidence between two deciding classifiers. This is done by a class-wise adaptive update process using past agreement coefficient values and the classifier decision likelihoods.

2.4.7 Challenges of fusion

Information fusion has its own challenges mainly rising from the differences between information streams - especially streams from different modalities. The following sections illustrate typical challenges faced in the fusion process.

Synchronization of multimodal data

The challenge of synchronization arises with the use of multiple separate recording devices in parallel for data acquisition. In order for the modalities to support each other in the recognition process their classification information needs to be in sync. Classic example of this synchronization is the use of a clapperboard in film production: the video track point of time, where the clapper board is clapped shut corresponds to the point of time in the audio tracks, where the clapping sound is heard.

In case of different capture rates the features might need to be interpolated in order to get into a common analysis granularity, depending on which kind of fusion methods are to be used. Synchronization is further complicated if some of the modalities have a nonuniform time representation such as shot segmentation, which is used with several motion features and with shot-based data in general [76].

With multiple recordings of the same modality various automatic similarity measure techniques can be used to find the time difference, for which the tracks are most similar, and synchronize the tracks based on this difference. These techniques are especially effective for audio recordings as they tend to be robust against position and orientation differences.

The cross-correlation of two one-dimensional signals is defined as the sequence of their sliding inner products, i.e. inner products with different delays between the two signals. The maximum of the cross-correlation function indicates the delay, at which the two signals are most similar. This can be used to synchronize two audio signals

with delay differences and additive noise due to capturing with different sensors at the same situation.

Processing time differences

In systems, where data is captured and analyzed online or processing resources are scarce, the processing time differences of the modalities can cause problems [2]. If some modality requires a lot of time or computational power to be processed, other information streams can suffer as their information goes to waste while waiting for the slow modality or their processing also gets delayed. Additional computational burden might also be introduced if additional synchronization is needed due to the processing delays. In these cases, more emphasis needs to be put on choosing efficient features and algorithms.

Confidence variation

The confidence of a multimedia stream represents its capability for accomplishing a multimedia analysis task. The most common form of incorporating confidence into a multimodal system is weighting of the separate multimedia streams before or during the fusion. The problem is that the confidence of a stream isn't constant and thus might be highly dependent on the data properties or contextual circumstances. For example, the confidence of the visual modality can drastically change in different lighting conditions. Confidence can also vary within a modality: color features usually work much better for color images than grayscale ones, whereas texture features typically work equally well in both cases. The weight assigning should thus be done dynamically based on the task and conditions. [2]

2.5 Parameter optimization

Finding the suitable combination of parameters may significantly affect the performance of parametric algorithms. It's impractical to go through all the parameter combinations with increased amount of parameters and the complex nonlinear relations between the parameter value combinations and the algorithm performance – especially in the case of continuous-valued parameters. Parameter optimization schemes try to find a balance between going through the parameter space to find the best combination of parameters and minimizing the time needed for the search.

2.5.1 Genetic algorithms

Genetic algorithms (GA) try to optimize the parameter space search by mimicing the principles of evolution. The parameters or more generally solutions to a

specific problem are discretized with enough precision and encoded on a simple chromosome-like data structure. A population of these chromosomes or genomes is randomly distributed into the solution space and their capacity of solving the problem is evaluated. Subsequently, a new generation of chromosomes is formed by simple operations inspired by natural selection and genetic variation. The chromosomes with higher evaluation scores are favored in reproducing with each other (e.g. forming new chromosomes by recombining their parts) and moving onto the next generation. Thus the potential solutions tend to converge to the optimal regions of the solution space. By mutating (randomly altering parts of) some of the chromosomes from time to time the search is offered the possibility to break free from local optima.

The population evaluation is done with a specific evaluation function, which calculates or approximates a solution to the problem. Approximation needs to be used in case calculating the actual problem solution takes a lot of time as the evaluation must be done separately for each chromosome at each generation. The search can be terminated when high enough performance is achieved or after a predefined amount of generations. More detailed description of genetic algorithms can be found in [89].

GAs have been applied to information fusion in [29]. The authors proposed one-dimensional GA-based weight search for late fusion, and more general multidimensional model to optimize multiple dimensions of the fusion process in parallel. Chen et al. in [18] also applied GAs for the fusion of fuzzy SVMs in a biomedical classification task.

2.6 Automatic visual lifelogging

The concept of digitally capturing daily activities and personal memories for later retrieval is known as lifelogging [11]. Automatic lifelogging aims at performing this task passively without user intervention. The idea was first outlined by Vannevar Bush as early as 1945 in his article *As We May Think* [83], where he describes the "memex", an automatic personal information storage and retrieval system. The idea was ahead of its time and had to wait decades for the advancement of information storage, data compression, and recording device technologies.

The information stored in a lifelog can be of diverse origin ranging from saving of communication activities, conversations, and positioning data to recording video or image sequences. The latter two cases are sometimes referred to as visual lifelogging to distinguish them from other types of lifelogs. Lately, as more and more people carry still image, audio, and video recording capable mobile devices with them practically all the time, the research on visual lifelogs has gained increased interest.

Visual lifelogging has sometimes been termed *sous*veillance as it's a form of inverse surveillance, where the surroundings are being monitored by the subject [23]. With

the combination of passive recording and remote storage, visual lifelogging could thus be used as a form of personal environment observation for safety purposes. As an example, a prospective mugger might be less likely to attack a person if the assaulter would know that his actions and characteristics or even identity would be captured. The lifelog data could also be used as a proof of what the user has or hasn't done [36]. In [54] Mann describes his views on the ethical and legal questions rising from this kind of subjective ubiquitous supervision, based on three decades of personal lifelogging experiments.

The concept of automatic visual lifelogging is intriguing as it provides means for capturing sudden moments of great importance, might allow basically perfect recollection of personal memories, or even perceiving things not noticed at the time of capture [36]. In a recent article by Kalnikaite et al. [40] different types of information stored in a lifelog are shown to promote different kind of acts of remembering, including inferencing of what happened rather than recalling the actual events. In another recent article the benefits of capturing everything one does and sees and replacing the human memory with an artificial system has been criticized [70]. Anyhow, even with selective partial capture the data amounts passively acquired with visual lifelogs would be so vast that efficient indexing and key frame selection methods are required for practical use of the data [22]. Another issue with lifelog recordings is the lack of predefined structure in contrast to broadcast video material. The lack of shots and cuts makes the segmentation and clustering of the continuous recordings a nontrivial task [49].

Content-based indexing provides the practical means to manage the massive and ever-growing lifelog archives, but additional metadata can be advantageous as well as suggested by Wang et al. in [86]. They believe that the best way to retrieve information from lifelogs is the combination of content-based analysis and attributes such as time and location. The two processes complement each other as attributes can refine content-based search, which in turn is able to retrieve similar unlabeled content even without any attribute information. In [41] additional methods for multisensory lifelog metadata creation are described.

3. MULTIMODAL VIDEO CONTEXT RECOGNITION FRAMEWORK

This chapter describes the details of a multimodal video context recognition framework implemented with C++ in Windows XP desktop environment. The framework is able to train and save SVM classifier models for arbitrary feature vectors from input files in a specific format. It also has the possibility to include saved SVM models and decision likelihoods of external systems from files. The framework fuses the internal and external decisions with weighted rule- or classification-based fusion. The correct classification rates of the individual decision systems and the fusers are reported based on evaluations on unseen test data. The term *expert* is hereby used to collectively refer to known internal classifiers and unknown external decision making systems.

The framework allows multiple iterations of processing with the input data randomly sampled to training, testing, and fuser testing sets at each iteration. The classification performance is calculated over the processing iterations. SVMs have been chosen as the base classifiers as well as for classification-based fusion due to their reported good performance in semantic video analysis tasks. However, the modular nature of the framework supports integration of supplemental classifiers and fusers by deriving them from the provided base classes. Although the framework was developed and tested for multimodal video context recognition task, it has been developed relatively independent of the application domain and could be used in other domains with slight modifications.

The executable files of the command line operated framework take less than 5 MB of storage space. Combined with the saved classification models of one processing iteration and the external acoustic likelihoods files used in the practical experimentations, the overall framework storage footprint is less than 25 MB. The framework would thus easily fit into a modern mobile device.

3.1 External frameworks

The multimodal video context recognition project started as a spinoff project from an audio context recognition project and experimented with combining the work of two research groups at Tampere University of Technology. The implementation utilizes external frameworks for preliminary data handling, such as the extraction of visual

features and the audio-based context recognition process. Algorithms and data structures from few freely available programming libraries have also been exploited.

3.1.1 MUVIS

MUVIS (**M**ultimedia **V**ideo **I**ndexing and **R**etrieval **S**ystem) [42] is a content-based multimedia indexing and retrieval framework developed by the MUVIS group at the Department of Signal Processing at Tampere University of Technology. Among several other functionalities the framework offers effortless multimedia database creation from diverse sources, easy modular integration of visual and acoustic feature extraction algorithms, and automatic feature extraction from video key frames.

MUVIS framework was utilized for visual feature extraction from the video database. The MUVIS database editor shown in Figure 3.1 was used to create a database from the videos as well as to extract a set of visual features from the video key frames. The feature files created by the database editor were read from file as input to the implemented framework. Further details of the feature extraction are given in section 3.2.2.

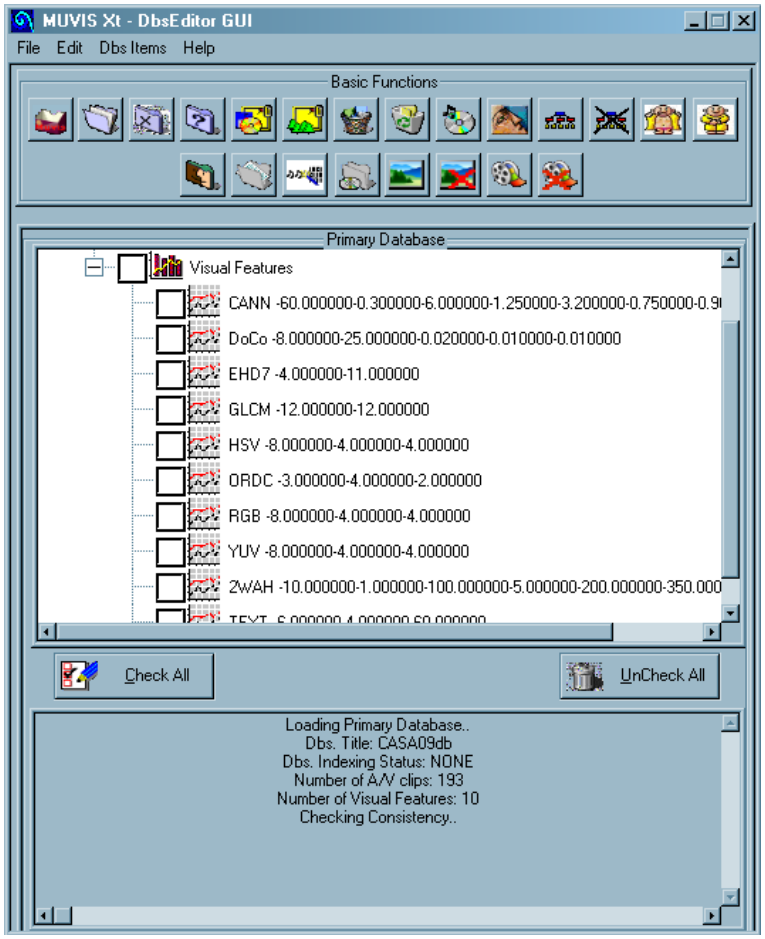

Figure 3.1: The main view of MUVIS DbsEditor application GUI showing a list of visual features extracted from a video database.

3.1.2 OpenCV

OpenCV (**Open** Source **C**omputer **V**ision) is a real-time computer vision library released under a BSD license [10]. It includes over 500 optimized algorithms for machine vision, signal processing, machine learning, and related tasks, with wrappers for C++, C, and Python.

OpenCV matrices and multidimensional arrays were used for the main data structures of the features and likelihoods inside the framework. These arrays allowed the efficient handling of subarrays, slices, rows, and columns of the data as the subparts could be referenced without any copying. Various operations readily implemented in the library were used to simplify and speed up the data handling. These included matrix operations, global extremum point locating, and data shuffling. OpenCV machine learning library was also used for initial SVM classifier implementations, but a more advanced SVM library was employed for the final SVM classifiers.

3.1.3 TUT audio context recognition framework

The audio context recognition framework [33] developed by Audio Research Group at the Department of Signal Processing at Tampere University of Technology in collaboration with Nokia Research Center Tampere conducts content-based audio context recognition based on audio event histograms models of the environments.

The audio framework was used to produce acoustic context likelihoods for the audio-visual database. These likelihoods along with synchronization information were fed as external expert decisions to the implemented framework.

3.1.4 LIBSVM

LIBSVM is a popular and profound SVM library capable of multiclass classification and soft likelihood output [13]. LIBSVM is available for free and commercial projects with a modified BSD license. The library is written in C++ and has interfaces to several languages and tools such as Java, MATLAB, Python, C#, and Perl.

LIBSVM was used in the framework for training, saving, loading, and evaluating the SVM models for the visual feature data. Likewise, the classification-based fusion was conducted with a LIBSVM SVM classifier.

3.1.5 GAlib

GAlib is a C++ library of genetic algorithm optimization tools. It is available free of charge and can be used for free or commercial purposes provided that the original author is credited [85].

GAlib was used for weight optimization in the rule-based fusion schemes. The GA

search was used to find the best performing weight combination for each rule-based fuser. GA search was also tried on the SVM parameter optimization, but simple grid search proved to be more applicable due to the simplicity of the task.

3.2 System input

The video context recognition was carried out by fusing recognition of aural and visual modalities. A video database along with high-quality audio recordings was gathered for the project. The database as well as the extracted video features are described in sections 3.2.1 and 3.2.2, respectively. Audio-based context recognition was done outside the implemented framework with the system described in [33] and the likelihoods of this analysis were imported into the framework as external expert information in the fusion phase.

Figure 3.2: Video frames of the 21 different contexts, from which data was recorded for the database.

3.2.1 Database

The database used for testing the framework consists of video and audio files recorded from 21 real-life contexts during summers of 2009 and 2010. The contexts were chosen to represent locations and situations, where people would typically record video with a mobile device. The used contexts are shown in Figure 3.2. Audio data was captured from 2 additional environments, but these contexts weren't used for the project due to the lack of video data. 2010 recordings also included a small set of clips with context changes during the recording. This subset was, however, not used in the framework development, so all the database clips consisted of material from a single fixed context. Table 3.1 shows the details of the recordings in the database.

The 2009 video recordings were done with a USB pen camera recording video

Table 3.1: Database recording details

Context	Recordings	Duration (hh:mm:ss)	Key frames	Year
Amusement park	9	28:40	337	2010
Basketball	7	2:09:01	175	2009
Beach	5	1:34:16	409	2009
Bus	1	27:39	124	2009
Café	10	31:50	389	2010
Family yard	8	25:28	309	2010
Football	16	50:55	624	2010
Hallway	10	1:43:37	1360	2009
Home	12	38:15	468	2010
Inside a train	11	35:01	425	2010
Nature	10	31:50	387	2010
Outdoor festival	10	31:49	380	2010
Outdoor market	13	41:23	489	2010
Party	6	19:04	232	2010
Pub or club	10	31:51	390	2010
Railway station	10	31:50	389	2010
Restaurant	10	1:34:28	707	2009
Shop	10	1:27:12	1272	2009
Sports	8	25:26	267	2010
Street	10	1:41:01	682	2009
Track'n'field	7	2:34:42	346	2009
Total	193	20:15:18	10161	–

at VGA resolution and mono audio sampled at 8 kHz. For the 2010 video recordings a mobile phone with video capabilities was used producing video at 176×144 resolution and audio at 8 kHz sampling rate. The separate high-quality audio was recorded both years uncompressed with bit depth of 24 bits and sampling rate of 44.1 kHz using a pocket recorder and binaural stereophonic microphones placed at the ears of the recording person. The audio and video recordings were started and stopped manually from the devices leading to desynchronization in the order of some seconds between the start and end points of the modalities. In rare cases the timing differences were more extreme due to battery failures or other problems. The video and audio recordings were synchronized temporally by maximizing the cross-correlation between each of the audio tracks of the video recordings and the correponding high-quality audio recordings. This was done outside the framework in MATLAB. The resulting offsets were verified by simultaneously listening to both the synchronized audio streams, and saved into a synchronization file given as an input to the framework.

The recordings ranged from 3 to 30 minutes of continuous recording at places or situations representing the context classes. Apart from the *bus* context all classes consisted of multiple takes. Depending on the context the recording was done at one place or on the move - whichever was natural behavior at any given context. Some contexts such as *hallway* contained recordings of both walking and sitting on a sofa. The recordings in 2009 were longer - usually from 10 to 30 minutes - and shot at one particular instance of a context class (i.e. always at the same restaurant, beach, etc.). However, the recording locations within the context were varied and no same routes were used on purpose between takes, when recording during walking. The 2009 videos contained a small graphic element showing the date and time on top of the video at the bottom right corner. Based on initial tests this didn't have a notable effect on discrimination between classes as the element was small and the framework used global low-level features.

All new set of contexts was used in the 2010 recordings, so there was no overlap on the data gathered with the different device setups. 2010 recordings were shorter in duration and represented the context classes in a more diverse manner with multiple locations within a context. The 2010 video data contained some unrepresentative material due to poor lighting conditions in some locations and handling the phone in an unobtrusive way when recording in certain public places. As these issues only affected a small set of videos from few contexts, and one of the project aims was to study how the audio and visual modalities complement each other in context recognition, the database was not filtered quality-wise before the feature extraction.

3.2.2 Visual features in the framework

Altogether six visual features were used in the audio-visual context recognition experimentations. Additional features were extracted from the database, but initial classification and fusion tests showed such low correct classification rates that the features were rejected from the final fusion tests and evaluations. The included features were HSV, RGB, and YUV color histograms [81], *gray-level co-occurrence matrix* (GLCM) [31], *ordinal co-occurrence matrix* (ORDC) [61], and *MPEG-7 edge histogram* (EHD7) [53].

The features were extracted globally from the video key frames and automatically saved by the MUVIS database editor application DbsEditor into feature files. The files included header information followed by the feature vectors of one video clip and one feature type. Based on the contexts of recording and key frame amount information from the feature files, a label file was also created for each video clip. As the videos contained continuous footage of one context, labeling all the key frames in the database was an effortless task and there was no need to consider unsupervised or semisupervised machine learning approaches.

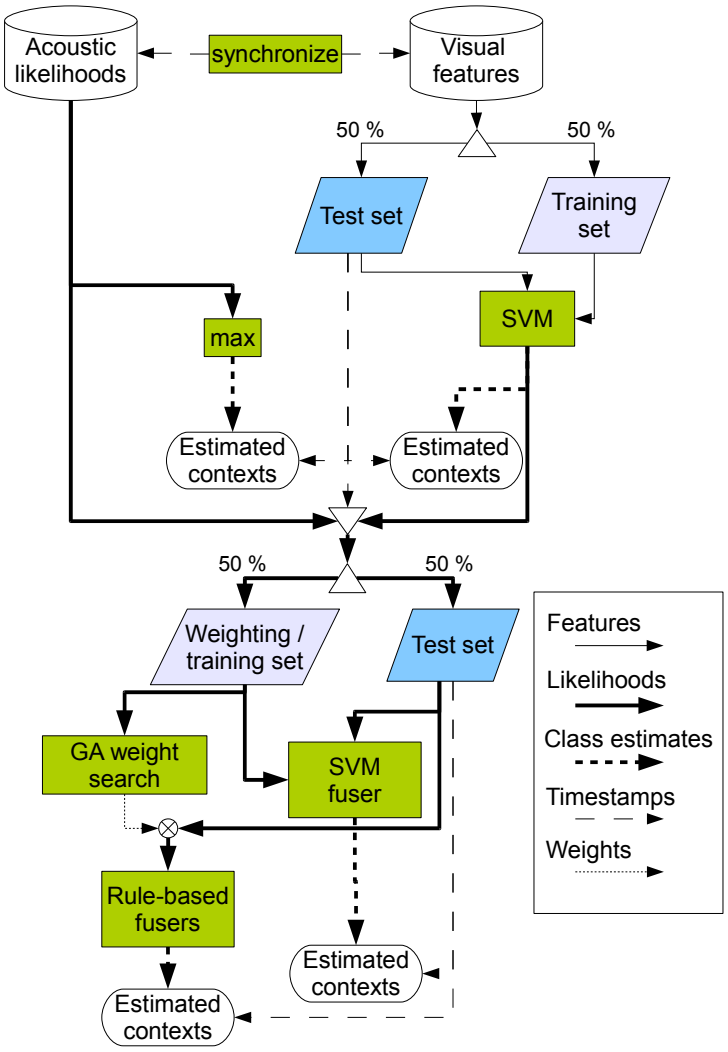

Figure 3.3: Flowchart showing the framework architecture.

3.3 Framework architecture

The framework architecture can roughly be divided into classification stage and fusion stage. In the classification stage input data is read from files and the individual classifiers trained on a subset of the data or pretrained classifier models loaded from files. Along with preparing the internal classifiers, external expert decisions for the data can also be loaded from files. Subsequently, the framework evaluates all the experts on an unseen test partition of the input data and reports their performance.

In the fusion stage, the context likelihoods of the experts on every test set sample are split into two subsets, one of which is used for rule-based fuser weighting and SVM fuser training, and the other for fusion evaluation. After weighting and training the fusers, the framework evaluates the fusion performance as correct classification rate by classifying the test set with each fuser. The split ratios in both classification

and fusion stages can be chosen by the user.

A flowchart describing the framework architecture is shown in Figure 3.3. The flowchart shows the framework as applied to the video context recognition task in the practical experimentations: the acoustic recognizer likelihoods are loaded as external expert decisions, SVMs trained on visual features used as internal classifiers, and 50 % data split ratios used in both the classification and the fusion stage. The whole processing starting from training or loading the individual classifiers can be iterated multiple times with different permutations of the input data chosen for training and testing, and the correct classification rate is calculated and averaged between iterations.

3.3.1 Input data handling

The framework reads in two types of data: feature vector sequences and detection likelihood sequences. Feature vector sequences along with their ground truth labels are used for training and testing of the individual experts. The detection likelihood sequences are used for fusing knowledge from external recognition systems with detectors inside the framework.

The features and the corresponding labels are fed to the system as a feature path list containing paths to the feature files in format described in section 3.2.2, and a label path list with paths to the label files of each database item (i.e. each video recording in the practical experimentations). The features and labels are matched, and the reading module checks that all the data samples have the same set of features extracted. After this, the data is read into a matrix, shuffled, and split into training and testing sets with a chosen split ratio.

External expert input files are handled similarly to the feature files. The system reads a path list of files containing estimated class occurrence likelihood sequences calculated for all classes from a single database item (i.e. an audio recording in the practical experimentations). Additionally, a synchronization difference file is also utilized for indicating the offset between the starting points of the feature sequences files and corresponding external likelihood sequences. With this synchronization information, the external likelihood sequences can be sampled from temporal positions corresponding to the timestamps of the input features.

3.3.2 Classification

The SVM classifiers of the framework can be trained from a subset of the input features or loaded as pretrained models from files that are saved separately for each data permutation in case of multiple iterations. Each feature type has its own classifier, which responds to an input feature vector with a vector of scores with each

element representing the likelihood of a certain class being present in the feature vector.

After training or loading the classifiers, class likelihoods are produced for the testing data, which is not used in the training phase. In case of using pretrained models, the data shuffling permutation and train/test splitting ratio have to be equal to the ones used, when the models were trained and saved. This is to ensure that the testing partition contains the same samples as when saving the model. Thus, no training data of the pretrained model is used for evaluating its performance. Accordingly, the external expert likelihoods are sampled at points corresponding to the test features and maximized over the classes to find the most prominent class of each test sample according to each external expert. Finally, the estimates of the internal and external experts are compared to the ground truth labels of the test data in order to get the individual recognition performances.

3.3.3 Decision fusion

The likelihoods produced by all experts are combined into an 3-dimensional likelihood structure, where the dimensions represent samples, experts, and classes. This likelihood array is again split into two parts over the sample dimension to get one set for rule-based fusion weighting as well as classification-based fusion training, and another for testing the fused performance. At this stage, the data is no more shuffled as it was already randomized in the classification phase.

Rule-based fusion

5 rule-based fusers are implemented and integrated into the framework: *Majority voting*, *sum of likelihoods*, *product of likelihoods*, *minimum likelihood*, and *maximum likelihood*. The experts are weighted separately for each fuser before the fusion. In Majority voting, each expert "votes" for its most likely class candidate, i.e. the crisp class label output, and votes are weighted and counted. The class getting the highest counter value is then declared as the fused output. Fusion by sum of likelihoods takes the sum of the weighted class likelihoods over the experts. Fusion output is the class, which gains the highest accumulated likelihood. Product of likelihoods works similarly to the sum of likelihoods with the exception of using the product operator instead of the sum operator. Minimum likelihood fusion finds the minimum likelihood among the weighted experts for each class, and finds the class having the highest minimum likelihood. This class is then chosen as the fusion output. Maximum likelihood fusion searches the highest weighted likelihood value among all the classes and experts and chooses the output class according to the maximum value.

For each fuser the experts are weighted with weights that are searched by a GA optimization procedure. The weight vector is used as the GA genome for the search. The weights are optimized by measuring the weighted fusion performance on the weighting data set for all the genomes at each generation and passing the best performing genomes onto the next generation via GA transform operations. Different weighting schemes are used for each fuser in order to be able to use the same weight value scale in all cases. Majority voting is weighted by multiplying the votes with the weight of the voting expert. As an example, if an expert has been assigned a weight value of 0.5 and it gives the highest likelihood score to the class *Café*, the counter of *Café* is incremented by 0.5. The influence of an expert is thus directly proportional to its weight value, if the weights of other experts are kept constant. With the four remaining rule-based fusers, specific weighting functions are used to weight all the likelihoods instead of the most likely crisp class output:

$$\tilde{l}_{ij} = w_i \cdot l_{ij} \mid 0 \leq w_i \leq a, a \in \mathbb{R}, \tag{3.1}$$

$$\tilde{l}_{ij} = l_{ij}^{w_i} \mid 0 \leq w_i \leq a, a \in \mathbb{R}, \tag{3.2}$$

$$\tilde{l}_{ij} = \begin{cases} (1 - w_i) \cdot \text{mean}(\check{\mathbf{l}}_{i(j)}) + w_i \cdot l_{ij} & \mid 0 \leq w_i \leq 1 \\ l_{ij}^{w_i} & \mid 1 < w_i \leq a, a \in \mathbb{R}, \end{cases} \tag{3.3}$$

and

$$\tilde{l}_{ij} = \begin{cases} (1 - w_i) \cdot \text{mean}(\check{\mathbf{l}}_{i(j)}) + w_i \cdot l_{ij} & \mid 0 \leq w_i \leq 1 \\ w_i \cdot l_{ij} & \mid 1 < w_i \leq a, a \in \mathbb{R}. \end{cases} \tag{3.4}$$

Equations 3.1, 3.2, 3.3, 3.4 show the weighting functions for fusion by sum of likelihoods, product of likelihoods, minimum likelihood, and maximum likelihood, respectively. In all the equations w_i represents the weight of ith expert, l_{ij} the likelihood of that expert for class j, $\tilde{l}_{ij}$ represents the weighted likelihood, where a is an upper limit for the weight values, and $\check{\mathbf{l}}_{i(j)}$ is the likelihood vector of the ith expert excluding the jth likelihood. Limiting the weight value is necessary for defining the search space for weight optimization.

In the sum of likelihoods fusion, the influence of an expert to the decision outcome is also directly proportional to its weight value, when the weights of other experts are assumed to be constant. The experts having zero valued weights have no effect on the fusion output. In product of likelihoods fusion, the likelihoods are raised to the power of the weight value. Hence, with weights close to zero the likelihoods are raised to the power of a value approaching zero, and the weighted likelihoods thus approach unity regardless of their value. Therefore they have lesser influence on the fusion. Accordingly, bigger weight values of an expert increase the dynamic range

and expand the differences between its likelihoods leading to increased influence in the multiplicative fusion process.

The weighting function of minimum likelihood fusion is defined piecewise around weight value of one. Weight values between zero and one are defined to bring the weighted likelihoods of the corresponding expert closer to the mean of the likelihoods of the remaining experts, as the weight approaches zero. As the weighted likelihoods approach the mean of the other experts, it's less probable that they are the minimum ones between all experts. Thus, with small weights an expert is less likely to have an impact on the fusion output. With weight values over one, the likelihoods are raised to the power of the weight value. As all the likelihood values have to be less than one and the weights are more than one by definition, the resulting weighted likelihoods will be smaller than before weighting. Thus, the higher the weight, the more probable it's to get the minimum weighted likelihood values from the corresponding expert.

The weighting function of maximum likelihood fusion is defined identically to minimum likelihood fusion with weight values between zero and one. Again, weights close to zero produce weighted likelihoods close to the mean of the likelihoods of other experts. Hence, it's less probable to find the maximum value among the likelihoods weighted with small weight values. With weight values above one, the likelihoods are multiplied with the weight value. Therefore, higher weights increase the probability of the corresponding expert having the maximum weighted likelihood over all experts. All in all, for each of the rule-based fusion schemes, small weight values give less emphasis to the expert and large values more.

Classification-based fusion

Classification-based fusion is carried out by concatenating the likelihood vectors of all the individual experts and feeding the concatenated likelihoods to an SVM. The SVM is trained with the classification-based fusion training partition of the likelihood values. Alternatively, the fusion SVM model can be read from a file. The SVM produces a vector with likelihoods for each class as a response to a concatenated likelihood vector. The most likely class is the fusion output.

4. EXPERIMENTAL RESULTS

The framework was evaluated by iterating the whole multimedia context recognition procedure 10 times for different expert combination and weight optimization choices. For the visual features, SVM classification the data was split in equal sized partitions, i.e. 5081 samples for training and 5080 for testing with random sampling at each iteration. The training sets were used to train the SVM classifiers for each feature, and testing sets for obtaining the likelihoods for fusion and testing the performance of the individual classifiers.

Acoustic recognizer likelihoods were requested from the timestamps of the test samples. Due to synchronization and duration differences between the audio and video recordings, acoustic likelihoods were available on average for 4614 of the requested timestamps. The remaining timestamps with no audio information were given uniform likelihoods for the acoustic expert to indicate the absence of knowledge from this modality. All the likelihoods were again divided into two equal sized sets. The first set was used for fuser training and weight optimization, and the second for evaluating the fused performance. Details of the database and features used for the experimentations are given in section 3.2.

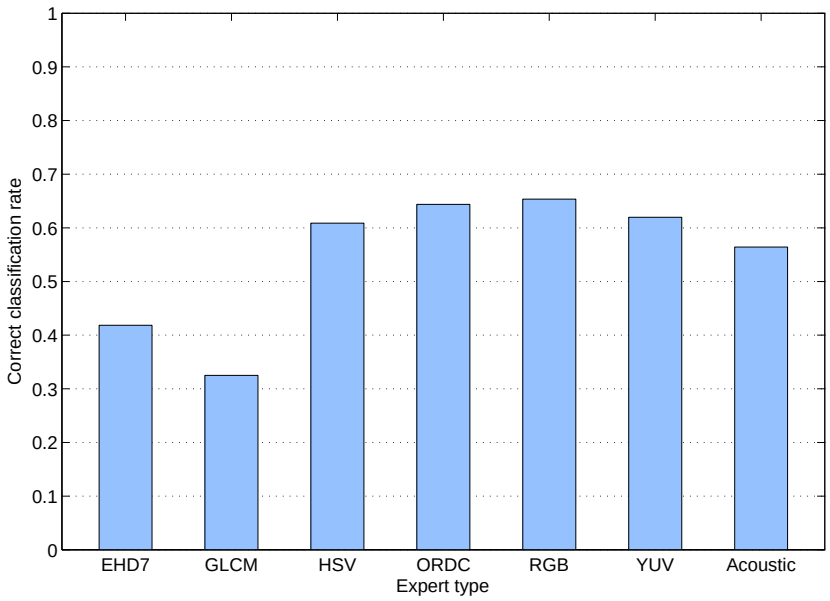

Figure 4.1: Correct classification rates of the individual experts on the database.

Table 4.1: Correct classification rates of the individual experts.

EHD7	GLCM	HSV	ORDC	RGB	YUV	Acoustic
0.418	0.325	0.609	0.644	0.653	0.620	0.564

4.1 Individual recognizer results

The individual experts classified the test samples with averaged correct classification rates ranging from 0.325 to 0.653 as shown in table 4.1. The best performance was achieved with the classifiers based on the RGB color histogram feature. Other color features gave rather similar performance. GLCM feature-based classifiers were responsible for the worst individual average performance. The second texture feature, ORDC, performed considerably better with almost double the rate of GLCM. The only edge feature, EHD7, was responsible for the second lowest correct classification rate. As clearly visible in Figure 4.1, there was a notable distinction between the four best performing visual experts and the two remaining ones. The acoustic expert produced correct classifications with a rate of 0.564.

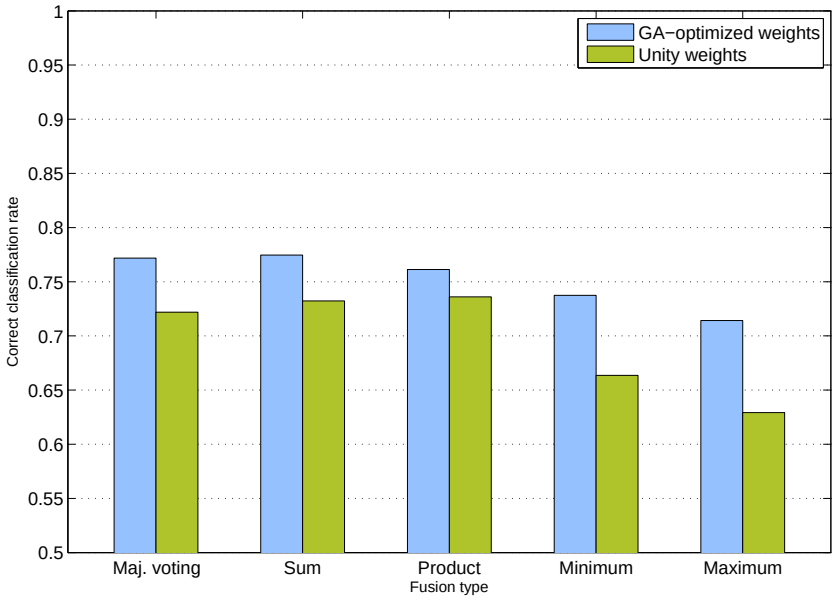

Figure 4.2: Correct classification rates with unity and GA-optimized weights for the three best performing visual and the audio-based expert.

4.2 Fusion results

The fusion was carried out by optimizing the weights of the rule-based fusers and training the SVM classifiers of each iteration from the first half of the likelihoods produced by the individual experts. Different genome and generation amounts were experimented with. The genome amount of 40 with 150 generations turned out to be satisfactory for the GA search to converge to a solution, since as high values as 200 genomes and 2000 generations didn't seem to result in better performing weights. Hence, the rule-based fusion evaluations were conducted with 40 genomes evolved over 150 generations. The GA weight search improved the performance with all the rule-based fusion methods compared to unity weights as shown in Figure 4.2.

For the non-piecewise defined rule-based fusers, i.e. majority voting, sum of likelihoods, and product of likelihoods, the scaling of all the weights of the fuser with a scalar value did not affect the performance. Thus, an experiment was made to fix one of the experts' weights, and let the GA search find the remaining weights freely. This was hypothesized to boost the weight optimization, as the search was constrained to one relative scale. In practice, however, no improvements were perceived and the fixing was not used in the final evaluations.

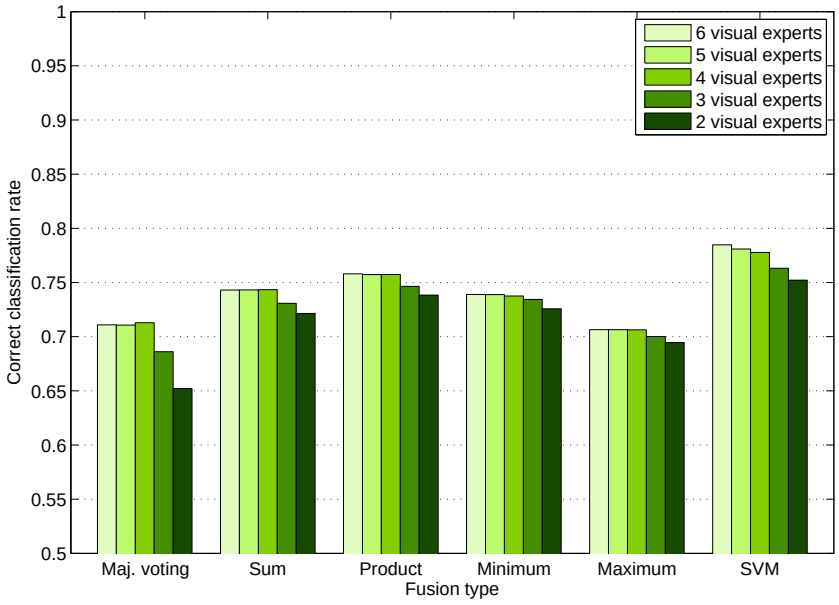

Figure 4.3: Correct classification rates of the fusion methods using different amounts of visual experts only.

The weight search space was defined to range from 0 to 20 in all expert dimensions. The upper limit was chosen to allow high enough weights for one expert to outweight the six others having unity weights. The optimized weight vectors were thus allowed to contain values from the mentioned range. The resulting weights were

then fed to the weighting functions defined in section 3.3.3 to produce the weighted likelihoods. 16 bits were allocated for the search range resulting in weight sensitivity of approximately $3.05 \cdot 10^{-4}$.

The weighting functions were designed to be able to ignore single experts by giving them zero weights. Nonetheless, the fusion process was tested with subsets of the features. The subsets were chosen by ordering the visual classifiers according to their individual correct classification rates and including varying amounts of the best performing individual experts. Additionally, the acoustic expert was excluded and included. When using only the visual modality, the weight search functioned well as the expert reduction didn't improve the results. Figure 4.3 shows the averaged correct classification rates for the different fusion methods using different amounts of best performing visual classifiers.

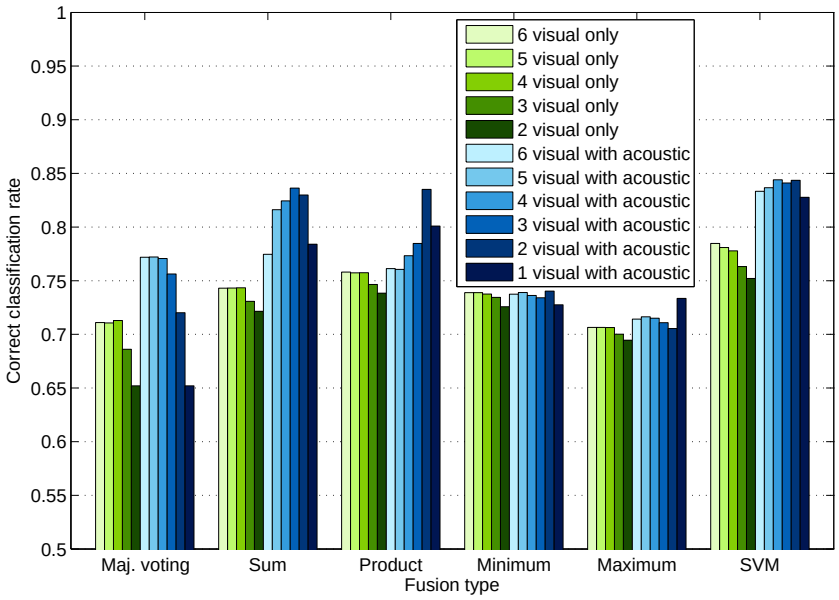

Figure 4.4: Correct classification rates of the fusion methods using different amounts of visual experts both with and without the acoustic expert.

Figure 4.4 shows the visual rates along with the corresponding multimodal results. With the inclusion of the audio modality two major observations were made: First of all, using all the experts the multimodal approach improved the performance in all fusion methods except for the minimum of likelihoods. The SVM fuser clearly outperformed all others in this case. Additionally, with three of the fusion methods, majority voting, sum of likelihoods, and SVM fusion, the improvements were far more significant than between the different amounts of visual experts. Secondly, in the multimodal approach the performance of all the fusion approaches was increased with the exclusion of some of the visual experts. Especially, the sum of

likelihoods, product of likelihoods, and maximum likelihood fusion methods experienced remarkable boosts in the correct classification rates. With the right choice of visual experts, sum and product of likelihoods were able to generate results comparable to the SVM fuser. Table 4.2 shows the correct classification rates of all the evaluated expert combinations highlighting the highest rate for each fuser.

Table 4.2: Correct classification rates of the different fusion methods.

Used experts	Maj. voting	Sum	Product	Minimum	Maximum	SVM
6 visual + acoustic	0.772	0.775	0.761	0.737	0.714	0.833
6 visual	0.711	0.743	0.758	0.739	0.706	0.785
5 visual + acoustic	**0.772**	0.816	0.761	0.739	0.716	0.837
5 visual	0.711	0.743	0.757	0.739	0.706	0.781
4 visual + acoustic	0.771	0.824	0.773	0.736	0.715	**0.844**
4 visual	0.713	0.743	0.757	0.738	0.706	0.778
3 visual + acoustic	0.756	**0.836**	0.785	0.734	0.711	0.841
3 visual	0.686	0.731	0.746	0.734	0.700	0.763
2 visual + acoustic	0.720	0.830	**0.835**	**0.740**	0.705	0.844
2 visual	0.652	0.721	0.738	0.726	0.695	0.752
1 visual + acoustic	0.652	0.784	0.801	0.728	**0.734**	0.828

The performance of the majority voting fusion approach deteriorated in both unimodal and multimodal cases with too drastic expert pruning. The majority voting fusion of two experts – both unimodal and multimodal cases – actually performed rather close to the best individual acoustic classifier. This performance drop was likely to be caused by the loss of decision resolution with less experts – especially, when voting from crisp decision labels. The same phenomenon was observable to a lesser degree also with the soft score based approaches except for maximum likelihood fusion, which actually produced its best correct classification rate with only one visual and the acoustic expert.

According to the experiments, minimum and maximum of likelihoods fusion methods were not suitable for combining the modalities. For instance, the sum of likelihoods and minimum likelihood fusers gave quite comparable results for the unimodal fusion of the visual experts, whereas after including the acoustic expert the former clearly outperformed the latter. This might be due to likelihood distribution differences between the modalities and the problem could thus be tackled to some extent by using some specific normalization scheme for the acoustic likelihoods. Anyhow, this input data dependent problem exists for the minimum and maximum likelihood fusers, while the other methods seemed not to be affected by it.

The SVM fusion proved to be the most robust method against the variation in the visual expert amount in the multimodal case. It was also responsible for the

Table 4.3: Average fusion training/weight search times using all 6 visual and the acoustic expert.

Fusion method	Processing time
Majority voting	26.7 s
Sum of likelihoods	31.7 s
Product of likelihoods	192 s
Minimum likelihood	217 s
Maximum likelihood	54.3 s
SVM	522 s

overall best correct classification rate of 0.844. This was achieved using 4 best performing visual experts along with the acoustic expert. With 3 digit precision the rate of the SVM approach with 2 best performing visual experts and the acoustic expert was also rounded up to 0.844. The drawback of SVM fusion is its computational complexity. As shown in table 4.3, SVM fusion took more than double the processing time compared to the slowest rule-based method, when fusing decisions of all available experts. Nonetheless, the rule-based fusion processing times are directly proportional to the chosen GA genome and generation amounts. Moreover, the fusion weight optimization and model training are offline processes.

4.3 Class-wise performance analysis

The fuser with the highest correct classification rate was further analyzed by calculating context-wise precision, recall, accuracy, and $F_{0.5}$ metrics from its classification results. All the metrics were averaged over the processing iterations. The context-wise binary metrics were calculated by considering the context under inspection as the positive class and all the other contexts as the negative class. Table 4.4 shows the metrics along with their average values over the contexts. The maximum metric values are highlighted in boldface.

The low recall of the bus context was likely due to having the smallest amount of samples out of all the contexts. Correspondingly, the relatively high amount of key frames with hallway and shop contexts decreased their accuracies. The contexts with the top $F_{0.5}$-measure values had in common that they had been recorded while sitting or standing in one place, and contained continuous distinctive sounds such as music for pub or club, crowd cheers for football, and wind gusts on trees and water splashing sounds for beach. Nevertheless, the track'n'field context got the second lowest $F_{0.5}$ score even though it had many similarities to football: crowd cheering sounds, as well as video of crowd, sky, red seats, and green field. Contexts recorded mainly while moving on foot or by bicycle were among the ones with the

lowest $F_{0.5}$ scores. This is understandable as they had the most heterogeneous sets of key frames, and depending on the context the auditory environment could change considerably between locations as well. The averages of the scores were weighted with the context sample amounts.

Table 4.4: 10 iteration average class-wise metrics of the fuser with the best average correct classification rate.

Context	Precision	Recall	Accuracy	$F_{0.5}$
Amusement park	0.808	0.761	0.986	0.783
Basketball	0.963	0.784	0.996	0.863
Beach	0.931	0.901	0.994	0.915
Bus	0.848	0.503	0.993	0.630
Café	0.714	0.801	0.982	0.755
Family yard	0.791	0.842	0.989	0.814
Football	**0.980**	**0.975**	**0.997**	**0.978**
Hallway	0.811	0.867	0.956	0.838
Home	0.905	0.935	0.993	0.920
Inside a train	0.954	0.959	0.996	0.957
Nature	0.903	0.838	0.989	0.869
Outdoor festival	0.841	0.827	0.986	0.834
Outdoor market	0.835	0.840	0.983	0.838
Party	0.899	0.899	0.995	0.899
Pub or club	0.913	0.888	0.992	0.900
Railway station	0.898	0.886	0.990	0.892
Restaurant	0.770	0.758	0.968	0.764
Shop	0.823	0.833	0.959	0.827
Sports	0.860	0.843	0.992	0.851
Street	0.726	0.738	0.964	0.732
Track'n'field	0.709	0.694	0.981	0.701
Average	0.841	0.841	0.978	0.839

4.4 Optimal fusion weights

The expert combinations with the highest correct classification rates for each rule-based fusion method were examined to find their optimal fixed weight values. The GA searched weights of the ten iteration rounds were analyzed. It was observed that majority voting, sum of likelihoods, and product of likelihoods fusion methods had weights with relatively low variation between iterations. The weight vectors of minimum and maximum likelihood fusers had clearly two distinctive groups. Hence, the optimal weights were estimated by averaging all the weights in the case of majority voting, sum of likelihoods, and product of likelihoods, and by taking the

local averages of the two groups of weights in the case of minimum and maximum likelihood fusers. Figure 4.5 shows the weight vectors of the iterations as well as the resulting averages. The weights of the fusers with piecewise defined weighting functions were not normalized, whereas the weights of the remaining fusion methods were normalized to sum to one before the averaging.

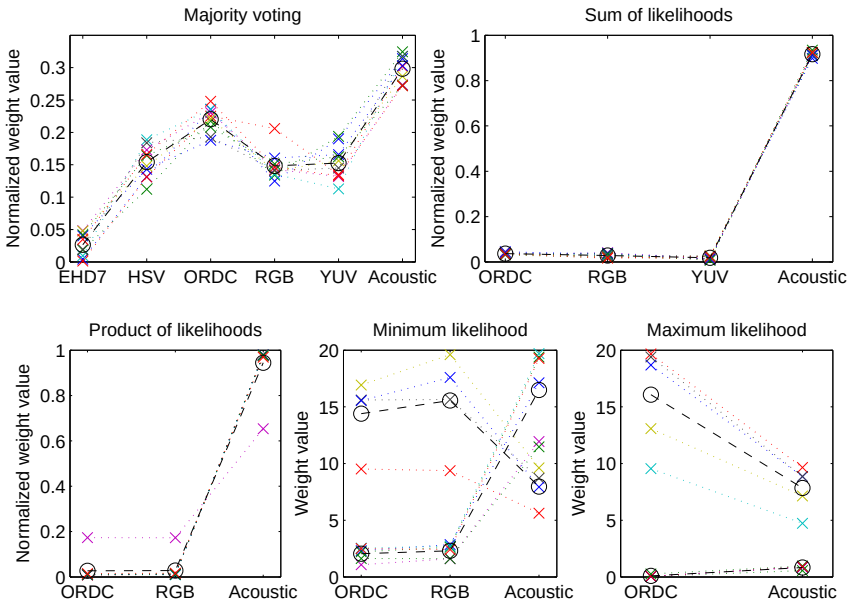

Figure 4.5: The weight distribution of the expert combinations with highest correct classification rate for each rule-based fusion method. Weight values are marked with × and average weights with ∘.

The average weights improved the best correct classification rates of majority voting, sum of likelihoods, minimum likelihood, and maximum likelihood fusion to 0.774, 0.837, 0.752, and 0.771, respectively. Product of likelihoods achieved a value of 0.817. Table 4.5 shows the optimum weights. For minimum and maximum likelihood fusion only the better performing of the two averaged weight vectors are given.

Table 4.5: Averaged optimal weight vectors for each rule-based fusion method.

Fusion method	EHD7	GLCM	HSV	ORDC	RGB	YUV	Acoustic
Majority voting	0.0265	–	0.154	0.220	0.148	0.153	0.298
Sum of likelihoods	–	–	–	0.0366	0.0283	0.0183	0.917
Product of likelihoods	–	–	–	0.0271	0.0284	–	0.945
Minimum likelihood	–	–	–	2.05	2.30	–	16.5
Maximum likelihood	–	–	–	–	0.0951	–	0.832

5. CONCLUSIONS

This thesis describes a multimodal video context recognition framework implemented in C++. The framework has been evaluated with real-world video and audio data recorded from 21 everyday audiovisual contexts. The contexts represent typical locations and situations, where people would record video with a mobile device. The framework uses altogether six global color, texture, and edge features extracted from video key frames to train SVM classifiers for context classification. The decisions of these classifiers are fused with an external audio-based context recognition system using various multimodal fusion methods. Genetic algorithms are utilized to search for optimal weights between the individual decision experts. Weighting functions for five rule-based fusion approaches have been presented. The framework has been designed in a flexible and modular way. Thus, adding further classifiers and fusion methods would be straightforward, and with minor adjustments the framework could be used for other classifier fusion tasks besides multimodal video context recognition. Although the implementation and experimentation was carried out in a desktop environment, the combined storage footprint of the implementation and the related classifier model files is small enough to easily fit into a modern mobile device.

In the performance evaluations the audio-based context recognition and the best visual-only fusion approach were able to achieve correct classification rates of 0.564 and 0.785, respectively. The correct classification rate of the best audiovisual fusion approach was 0.844. This was achieved by excluding the two visual classifiers with the lowest individual correct classification rates, combining the context likelihoods of the remaining experts, and classifying them with an SVM classifier. The multimodal fusion clearly improves the performance of the audio-based context recognizer, and also outperforms the visual-only approach. This shows that for the task of video context recognition, it is advantageous to use information from multiple modalities. Although the classification-based fuser outperformed all the simple rule-based fusion methods, the best rule-based fusion approach was able to achieve correct classification rate of 0.837, which is rather close to the highest performing fuser. Hence, it would be recommendable to also further study particular rule-based fusion methods in future work.

It was observed from the analysis of the optimal weight values of the best per-

forming rule-based fusers, that the audio modality is weighted with considerably higher weights compared to the visual classifiers. This might be due to different likelihood distributions between the modalities. However, the audio-based recognition system produces likelihoods with less noise. In other words, it produces high likelihood values only when being certain of the output context. Thus, even with relatively low weights the visual system gets to decide the classification outcome, if the acoustic expert is uncertain.

A few possible points of improvement should be considered for future work. The utilized key frame based sampling is non-optimal for audiovisual data. While the average key frame sampling frequency is higher in contexts and recordings with more variation in the video content – e.g. recordings made when moving – the audio modality is not taken into account in the data sampling. Time points with distinctive audio information may thus be left outside the evaluation. The external audio context recognizer provides context likelihoods with uniform sampling. Therefore, more suitable data sampling methods, such as key frame interpolation, should be examined in later work. Additionally, the data imbalance between contexts with more frequently sampled data and those with less samples should be taken into account in the processing.

The multiclassification approach adopted in the framework has the drawback that with increased amount of contexts each classifier has to distinguish between more and more classes. Moreover, with each context addition, all the classifiers need to be trained anew. Binary context-wise detectors would allow optimizing the features for each context. When adding a new context to the system, in theory only the corresponding detector would need to be trained, which would increase the framework scalability and flexibility. Anyhow, in practice with large enough sets of new data and contexts, the original detectors would have to be retrained with additional negative samples drawn from the added contexts for optimal discriminatory performance. Incremental learning approaches could be studied to overcome this problem. Transferring from multiclassifiers to binary classifiers might highly increase the overall amount of classifiers needed to train. However, binary classifiers should be much simpler and faster to train than multiclassifiers – especially if the class amount was substantially high.

Inter-concept relations and temporal dependencies have not been explicitly studied, since the contexts used in the practical experimentations are mainly mutually exclusive in nature, and no recordings with context changes were used in the database. In future work with more versatile data, it could be beneficial also to examine these context dependence related matters. Additionally, more intelligent local, salient point, shape, and motion features should also be investigated in order to carry out more comprehensive analysis of the typical recurrent objects, views, and

events at a certain context. This concept-based approach would add to the robustness of the context recognition by deepening the general semantic understanding of the framework. Moreover, investigating inter-concept relations might be more suitable for this type of analysis of mutually non-exclusive concepts.

Further weight search methods should be investigated for rule-based fusion as the GA weight optimization between all the experts was not able to find the solutions found by explicitly excluding some of the experts. The problem was observed especially in the multimodal fusion case. Significant differences were also noted in the ability of different fusers to utilize diverse information between modalities. Thus, the diversity between the different modalities could be estimated with diversity measures, and this information utilized for the optimization of the fusion. Additionally, with per-context detectors, context-wise confidence estimation could be used to choose optimal modalities, features, and algorithms for each context class.

As modern mobile devices generally contain a wide range of sensors besides cameras and microphones, ranging from wireless network sensors to positioning systems and accelerometers, the utilization of additional modalities should also be examined. Constantly more widely used wireless broadband connections enable additional information streams, such as weather measurements, and location tagged photographs. Video context recognition could benefit from these supplemental modalities by evaluating the confidence of all streams for each context, and using the streams accordingly for aiding the recognition task.

REFERENCES

[1] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient retrieval of life log based on context and content. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 22–31, New York, NY, USA, 2004. ACM.

[2] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, pages 1–35, 2010. 10.1007/s00530-010-0182-0.

[3] S. Ayache, G. Quénot, and J. Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *ECIR'07: Proceedings of the 29th European conference on IR research*, pages 494–504, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *Multimedia, IEEE Transactions on*, 4(1):68 –75, mar. 2002.

[5] S. Benini, L. Canini, P. Migliorati, and R. Leonardi. Multimodal space for rushes representation and retrieval. In *CBMI '09: Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pages 50–55, Washington, DC, USA, 2009. IEEE Computer Society.

[6] R. Benmokhtar and B. Huet. Perplexity-based evidential neural network classifier fusion using mpeg-7 low-level visual features. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 336–341, New York, NY, USA, 2008. ACM.

[7] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 368–374, Cambridge, MA, USA, 1999. MIT Press.

[8] M. Blighe and N. E. O'Connor. Myplaces: detecting important settings in a visual diary. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 195–204, New York, NY, USA, 2008. ACM.

[9] M. Blum, A. Pentland, and G. Troster. Insense: Interest-based life logging. *Multimedia, IEEE Transactions on*, 13(4):40–48, 2006.

[10] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[11] D. Byrne, A. R. Doherty, C. G. M. Snoek, G. G. Jones, and A. F. Smeaton. Validating the detection of everyday concepts in visual lifelogs. In *SAMT '08: Proceedings of the 3rd International Conference on Semantic and Digital Media Technologies*, pages 15–30, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] J. Calic, N Campbell, S. Dasiopoulou, and Y Kompatsiaris. An overview of multimodal video representation for semantic analysis. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies (EWIMT 2005), IEE*, 2005.

[13] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264, New York, NY, USA, 2007. ACM.

[15] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna. Semantic event detection via multimodal data mining. In *IEEE Signal Processing Magazine*, pages 38–46, 2006.

[16] M. Chen, C. Zhang, and S.-C. Chen. Semantic event extraction using neural network ensembles. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 575–580, Washington, DC, USA, 2007. IEEE Computer Society.

[17] S.-C. Chen, M.-L. Shyu, and M. Chen. An effective multi-concept classifier for video streams. In *ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 80–87, Washington, DC, USA, 2008. IEEE Computer Society.

[18] X. Chen, R. Harrison, and Y.-Q. Zhang. Genetic fuzzy fusion of svm classifiers for biomedical data. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pages 654 –659 Vol.1, sep. 2005.

[19] N. Dimitrova, L. Agnihotri, and G. Wei. Video classification based on hmm using text and faces. In *In European Signal Processing Conference*, 2000.

[20] Y. Ding and G. Fan. Sports video mining via multichannel segmental hidden markov models. *Multimedia, IEEE Transactions on*, 11(7):1301–1309, 2009.

[21] C. Djeraba. Content-based multimedia indexing and retrieval. *Multimedia, IEEE Transactions on*, 9(2):18 –22, apr. 2002.

[22] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. F. Jones, and M. Hughes. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 259–268, New York, NY, USA, 2008. ACM.

[23] A. R. Doherty, C. Ó Conaire, M. Blighe, A. F. Smeaton, and N. E. O'Connor. Combining image descriptors to effectively retrieve events from visual lifelogs. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 10–17, New York, NY, USA, 2008. ACM.

[24] A. R. Doherty, A. F. Smeaton, K. Lee, and D. P. W. Ellis. Multimodal segmentation of lifelog data. In *In Proc. of RIAO 2007*, 2007.

[25] L.-Y. Duan, M. Xu, X.-D. Yu, and Q. Tian. A unified framework for semantic shot classification in sports videos. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 419–420, New York, NY, USA, 2002. ACM.

[26] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[27] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *Proc. IEEE Intl. Conf. Multimedia and Expo*, pages 881–884, 2006.

[28] J. Fan, H. Luo, J. Xiao, and L. Wu. Semantic video classification and feature subset selection under context and concept uncertainty. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 192–201, New York, NY, USA, 2004. ACM.

[29] B. Gabrys and D. Ruta. Genetic algorithms in classifier fusion. *Appl. Soft Comput.*, 6(4):337–347, 2006.

[30] F. Gouyon, S. Dixon, and G. Widmer. Evaluating low-level features for beat classification and tracking. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1309 –IV–1312, apr. 2007.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610 –621, nov. 1973.

[32] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.

[33] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen. Audio context recognition using audio event histograms. In *Proc. 18th European Signal Processing Conference (EUSIPCO)*, 2010.

[34] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Trans. Inf. Syst.*, 27(3):1–29, 2009.

[35] S. C. H. Hoi and M. R. Lyu. A multimodal and multilevel ranking scheme for large-scale video retrieval. *Multimedia, IEEE Transactions on*, 10(4):607–619, June 2008.

[36] T. Hori and K. Aizawa. Context-based video retrieval system for the life-log applications. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 31–38, New York, NY, USA, 2003. ACM.

[37] D. P. Huijsmans and N. Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):245–251, 2005.

[38] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, pages 5–14, New York, NY, USA, 2009. ACM.

[39] E. Kaasinen. User needs for location-aware mobile services. *Personal Ubiquitous Comput.*, 7(1):70–79, 2003.

[40] V. Kalnikaite, A. Sellen, S. Whittaker, and D. Kirk. Now let me see where i was: understanding how lifelogs mediate memory. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 2045–2054, New York, NY, USA, 2010. ACM.

[41] I.-J. Kim, S. C. Ahn, H. Ko, and H.-G. Kim. Automatic lifelog media annotation based on heterogeneous sensor fusion. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 703 –708, aug. 2008.

[42] S. Kiranyaz, K. Caglar, E. Guldogan, O. Guldogan, and M. Gabbouj. Muvis: a content-based multimedia indexing and retrieval framework. In *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*, volume 1, pages 1 – 8 vol.1, jul. 2003.

[43] L. I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):281–286, 2002.

[44] N. Lazarevic-McManus, J. Renno, and G. A. Jones. Performance evaluation in visual surveillance using the f-measure. In *VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 45–52, New York, NY, USA, 2006. ACM.

[45] M. L. Lee and A. K. Dey. Wearable experience capture for episodic memory support. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 107 –108, sep. 2008.

[46] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[47] Z. Li and Y.-P. Tan. Event detection using multimodal feature analysis. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 3845 – 3848 Vol. 4, 23-26 2005.

[48] L. Lin and M.-L. Shyu. Mining high-level features from video using associations and correlations. In *ICSC '09: Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pages 137–144, Washington, DC, USA, 2009. IEEE Computer Society.

[49] W.-H. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *IS & T/SPIE Symposium on Electronic Imaging*, 2006.

[50] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *Multimedia, IEEE Transactions on*, pages 240–251, 2008.

[51] X. Liu, M. Corner, and P. Shenoy. Seva: Sensor-enhanced video annotation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(3):1–26, 2009.

[52] Y. Liu, D. Zhang, G. Lu, and A.-H. Tan. Integrating semantic templates with decision tree for image semantic learning. In Tat-Jen Cham, Jianfei Cai, Chitra Dorai, Deepu Rajan, Tat-Seng Chua, and Liang-Tien Chia, editors,

Advances in Multimedia Modeling, volume 4352 of *Lecture Notes in Computer Science*, pages 185–195. Springer Berlin / Heidelberg, 2006. 10.1007/978-3-540-69429-8_19.

[53] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703 –715, jun. 2001.

[54] S. Mann. Sousveillance and cyborglogs: a 30-year empirical voyage through ethical, legal, and policy issues. *Presence: Teleoper. Virtual Environ.*, 14(6):625–646, 2005.

[55] K. Z. Mao. Feature subset selection for support vector machines through discriminative function pruning analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(1):60–67, 2004.

[56] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(8):1778 – 1790, aug. 2004.

[57] H. Mizuno, K. Sasaki, and H. Hosaka. Indoor-outdoor positioning and lifelog experiment with mobile phones. In *WMISI '07: Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*, pages 55–57, New York, NY, USA, 2007. ACM.

[58] M. Mühling, R. Ewerth, T. Stadelmann, B. Freisleben, R. Weber, and K. Mathiak. Semantic video analysis for psychological research on violence in computer games. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 611–618, New York, NY, USA, 2007. ACM.

[59] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Combining multimodal and temporal contextual information for semantic video analysis. In *ICIP'09: Proceedings of the 16th IEEE international conference on Image processing*, pages 4269–4272, Piscataway, NJ, USA, 2009. IEEE Press.

[60] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 17(3):423–435, 2009.

[61] M. Partio, B. Cramariuc, and M. Gabbouj. An ordinal co-occurrence matrix framework for texture retrieval. *J. Image Video Process.*, 2007(1):1–1, 2007.

[62] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang. Correlative multilabel video annotation with temporal kernels. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(1):1–27, 2008.

[63] J. Qin, Y. Li, and W. Sun. A semisupervised support vector machines algorithm for bci systems. *Intell. Neuroscience*, 2007:12–12, 2007.

[64] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han. Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, pages 721–724, New York, NY, USA, 2009. ACM.

[65] R. Ranawana and V. Palade. Multi-classifier systems: Review and a roadmap for developers. *Int. J. Hybrid Intell. Syst.*, 3(1):35–61, 2006.

[66] M. I. Ribeiro. Kalman and extended kalman filters: Concept, derivation and properties. Technical report, Institute for Systems and Robotics, Lisboa, 2004.

[67] L. Rokach. Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1-2):1–39, 2010.

[68] D. Ruta and B. Gabrys. An overview of classifier fusion methods, 2000.

[69] F. Schaffalitzky and A. Zisserman. Automated location matching in movies. *Comput. Vis. Image Underst.*, 92(2-3):236–264, 2003.

[70] A. J. Sellen and S. Whittaker. Beyond total capture: a constructive critique of lifelogging. *Commun. ACM*, 53(5):70–77, 2010.

[71] S. T. Shivappa, M. M. Trivedi, and B. D. Rao. Audio-visual information fusion in human computer interfaces and intelligent environments: A survey. In *Proceedings of the IEEE, 2010*, 2010.

[72] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *Multimedia, IEEE Transactions on*, 10(2):252–259, 2008.

[73] A. F. Smeaton, B. Lehane, N. E. O'Connor, C. Brady, and G. Craig. Automatically selecting shots for action movie trailers. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 231–238, New York, NY, USA, 2006. ACM.

[74] C. G. M. Snoek, O. De Rooij, B. Huurnink, J. C. Van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. Van Liempt, R. Van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. De Rijke, J. M. Geusebroek,

T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma. The mediamill trecvid 2008 semantic video search engine draft notebook paper.

[75] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *Multimedia, IEEE Transactions on*, 9(5):975 –986, aug. 2007.

[76] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2(4):215–322, 2008.

[77] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *Multimedia, IEEE Transactions on*, 9:280–292, 2007.

[78] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM.

[79] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM.

[80] Y. Song, X.-S. Hua, L.-R. Dai, M. Wang, and R.-H. Wang. An automatic video semantic annotation scheme based on combination of complementary predictors. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V –V, may. 2006.

[81] M. J. Swain and D. H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, 1991.

[82] B. L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J. R. Smith. Normalized classifier fusion for semantic visual concept detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II – 535–8 vol.3, sep. 2003.

[83] B. Vannevar. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[84] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[85] M. Wall. GAlib: A C++ library of genetic algorithm components. *Mechanical Engineering Department, Massachusetts Institute of Technology*, 1996.

[86] Z. Wang, M. D. Hoffman, P. R. Cook, and K. Li. Vferret: Content-based similarity search tool for continuous archived video. In *in CARPE Third ACM workshop on Capture, Archival and Retrieval of Personal Experiences*, 2006.

[87] S. Wei, Y. Zhao, Z. Zhu, and N. Liu. Multimodal fusion for video search reranking. *IEEE Transactions on Knowledge and Data Engineering*, 22:1191–1199, 2010.

[88] M.-F. Weng and Y.-Y. Chuang. Multi-cue fusion for semantic video indexing. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 71–80, New York, NY, USA, 2008. ACM.

[89] D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4:65–85, 1994.

[90] K. Wickramaratna, M. Chen, S.-C. Chen, and M.-L. Shyu. Neural network based framework for goal event detection in soccer videos. In *ISM '05: Proceedings of the Seventh IEEE International Symposium on Multimedia*, pages 21–28, Washington, DC, USA, 2005. IEEE Computer Society.

[91] F. Wu, Y. Liu, and Y. Zhuang. Tensor-based transductive learning for multimodality video semantic concept detection. *Multimedia, IEEE Transactions on*, 11(5):868–878, 2009.

[92] G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced dataset learning. In *In ICML 2003 Workshop on Learning from Imbalanced Data Sets*, pages 49–56, 2003.

[93] Z. Wu, L. Cai, and H. Meng. M.h.: Multi-level fusion of audio and visual features for speaker identification. In *In: Proc. Int. Conf. Biometrics, LNCS 3832*, pages 493–499, 2006.

[94] L. Xie and S.-F. Chang. Pattern mining in visual concept streams. In *IEEE International Conference on Multimedia and Expo (ICME 06)*, Toronto, Canada, 2006.

[95] H. Xu and T.-S. Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):44–67, 2006.

[96] M. Xu, S. Luo, J. S. Jin, and M. Park. Affective content analysis by mid-level representation in multiple modalities. In *ICIMCS '09: Proceedings of the First International Conference on Internet Multimedia Computing and Service*, pages 201–207, New York, NY, USA, 2009. ACM.

[97] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical report, Columbia University ADVENT, 2007.

[98] L. Yang, J. Liu, X. Yang, and X.-S. Hua. Multi-modality web video categorization. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 265–274, New York, NY, USA, 2007. ACM.

[99] X. Zhang, Y.-C. Song, J. Cao, Y.-D. Zhang, and J.-T. Li. Large scale incremental web video categorization. In *WSMC '09: Proceedings of the 1st workshop on Web-scale multimedia corpus*, pages 33–40, New York, NY, USA, 2009. ACM.

[100] Y.-F. Zhang, C. Xu, H. Lu, and Y.-M. Huang. Character identification in feature-length films using global face-name matching. *Multimedia, IEEE Transactions on*, 11(7):1276–1288, 2009.

[101] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[102] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans. on Knowl. and Data Eng.*, 17(5):665–677, 2005.